

sasCIF: an extension of core Crystallographic Information File for SAS

M. Malfois and D.I. Svergun

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site provided that this cover page is retained. Republication of this article or its storage in electronic databases or the like is not permitted without prior permission in writing from the IUCr.

sasCIF: an extension of core Crystallographic Information File for SAS

Marc Malfois^a and Dmitri I. Svergun^{ab*}

^aEMBL c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

^bInstitute of Crystallography, Russian Academy of Sciences, Leninsky

prospekt 59, 117333 Moscow, Russia

Email: Svergun@EMBL-Hamburg.de

Data acquisition packages developed at different small angle scattering facilities use different formats both for raw and processed data storage. To facilitate the data exchange between laboratories, a consensus in the small angle scattering community has been reached on an ASCII format for one-dimensional data which includes a self-describing header containing relevant information about the sample and instrumental conditions followed by raw or reduced data in a tabular form. This format called sasCIF was implemented as an extension of core CIF (Crystallographic Information File) dictionary.

1. Introduction

Growing number of users and user groups especially those at large scale facilities make the problem of small angle scattering data portability and exchange more and more important. Due to differences in data acquisition software and hardware, local raw data file formats are different. Moreover, the reduced data are often stored in different formats, too. The need to compare results obtained at different facilities leads to a multiplicity of software packages for analysing data and most of them are able to read/write a single fixed data format only. It is always possible to write a program converting from one format into another; however, with intensifying X-ray and neutron scattering data exchange, the users might spend more time writing subroutines than analysing the data. Moreover, useful information about the experiment can be lost in the converting process.

The general awareness of the problem (canSAS I at Grenoble, February 1998 and canSAS II at Brookhaven laboratory, May 1999) has led to a decision to acquire a universal format that would be also important to make SAS more accessible to new users and to improve the understanding of the measurement process. The obvious requirements for this format are: it must be self-describing, flexible, extensible and portable. The self-description means full definition of all names and concepts used, flexibility permits to use only the names describing the current experiment, extensibility means that addition of new items must not disturb the existing files and portability is the accessibility of data items independently of their point of origin (e.g. on different computers).

Up to now, several approaches were explored to facilitate the data exchange. The approach presented here is a proposed extension of the Crystallographic Information File (CIF) data representation used for describing small molecule structure and associated diffraction experiments (Hall *et al.*, 1991). This extension is called the small angle scattering Crystallographic Information File (sasCIF) and is intended to store one-dimensional (e.g. after radial or sectorial average) experimental data and associated parameters describing the small angle scattering experiment. As the CIF format is sponsored by

the International Union of Crystallography (IUCr), a wide community is using it, and useful conversion software routines are available from the public domain. The CIF format is an ASCII format, it is flexible, extensible, portable and each value is related to a keyword, that is, it fulfils the requirements of a universal format. Some extensions are already available: a pdCIF dictionary for the powder diffraction and a mmCIF dictionary for the macromolecular structures (Bourne *et al.*, 1997). Development of other extensions e.g. an imgCIF dictionary for data collected with a two-dimensional detector is in progress.

2. General rules

The sasCIF dictionary is an extension of the CIF dictionary and for this reason, some rules must be applied to the sasCIF dictionary to be compliant with the core CIF dictionary. Each keyword defining a value is constructed in the form `_sas_category.extension`. The prefix `_sas_` specifies that the keyword belongs to the sasCIF dictionary extension. Data items describing the same part of the experiment (for instance, the data items relevant to describe the sample) are grouped in the same category. CIF permits to use loops for repeating items in the same category in which the values are separated by a white space or a new line. If the value is a character string on a single line, it can be delimited by single (') or double (") quotes so that the white spaces are not considered as separators. The values extended beyond a single line are enclosed by semicolons (;). The first semicolon is the first character of the line where the text block starts and the second one is the first character of the line following the last line of text. The record length is restricted to 80 characters and only ASCII characters are allowed. The comments are preceded by a hash (#) and are terminated by a new line. Data values which are unknown or undefined are represented by a question mark (?) and a period (.) respectively. At present, only a single level of loop is permitted. CIF (and the sasCIF) dictionary contains the textual descriptions, attributes of the data items like units, ranges of values (if any) and indicators whether the data item is mandatory. Which data items will be mandatory and which units will be used to describe small angle scattering experiments adequately has still to be decided by community.

The sasCIF dictionary is built with the Dictionary Description Language (DDL) version 2.2.1 created for the mmCIF (Westbrook *et al.*, 1995). With this version of DDL, even the items defined in the core CIF dictionary must be written in the sasCIF dictionary, i.e. the latter contains all the keywords relevant for the SAS experiment.

3. Structure of a sasCIF file.

A sasCIF file can be described by categories included in the CIF dictionary and by four related categories necessary to describe a small angle experiment.

The SAS_BEAM category gives information on the experiment geometry. In this category, item names defining the shape and the size of the beam or the distance between the elements of the instrument are described. The SAS_DETC category provides description of the detector (the number of pixels, orientation, beamstop shape and sizes etc...). The sample environment is described in the SAS_SAMPLE category that contains physical and chemical details about the sample consisting of the specimen (e.g. particles) and the matrix (e.g. solvent). The scattering data themselves are described by the SAS_INTENSITY category related

Table 1

Structure of the sasCIF file containing more than one scattering experiment. The pointers are used in this example to indicate that the sample 1 and the matrix 1 were measured with a long camera length and a linear detector while the sample 2 and the matrix 2 at a short camera length and a quadrant detector.

```
Data_TEST1                # Identifier of the first data block.
_sas_beam.id              Long_camera
# Beam geometry description
_sas_detc.id              Linear_detector
_sas_detc.beam_id         Long_camera
# Detector description. All items in this category will refer to the items describing the
# beam geometry called "Long_camera"
_sas_sample.id            Sample_1
_sas_sample.beam_id       Long_camera
_sas_sample.detc_id       Linear_detector
# First Sample description. All items in this category will refer to the items describing
# the beam geometry called "Long_camera" and the detector called "Linear_detector"
_sas_intensity.sample_id  Sample_1
# Data values for sample_1
_sas_sample.id            Matrix_1
_sas_sample.beam_id       Long_camera
_sas_sample.detc_id       Linear_detector
# First matrix description. All items in this category will refer to the items describing
# the beam geometry called "Long_camera" and the the detector called
# "Linear_detector"
_sas_intensity.sample_id  Matrix_1
# Data values for matrix_1
_sas_beam.id              Short_camera
# Beam geometry description.
_sas_detc.id              Quadrant_detector
_sas_detc.beam_id         Short_camera
# Detector description. All items in this category will refer to the items describing the
# beam geometry called "Short_camera"
_sas_sample.id            Sample_2
_sas_sample.beam_id       Short_camera
_sas_sample.detc_id       Quadrant_detector
# Second sample description. All items in this category will refer to the items
# describing the beam geometry called "Short_camera" and the detector called
# "Quadrant_detector"
_sas_intensity.sample_id  Sample_2
# Data values for sample_2
_sas_sample.id            Matrix_2
_sas_sample.beam_id       Short_camera
_sas_sample.detc_id       Quadrant_detector
# Second matrix description. All items in this category will refer to the items
# describing the beam geometry called "Short_camera" and the detector called
# "Quadrant_detector"
_sas_intensity.sample_id  Matrix_2
# Data values for matrix_2
# End of file
```

to the SAS_SAMPLE category through an identifier. As the sasCIF format permits to store several data sets in one file, it is necessary to relate the intensity category to the sample category, which is in turn related to the beam and detector categories identified by a character string as shown in Table 1. This feature is useful when several experiments are merged in a single sasCIF file. The scattering

intensity is written in a tabular form. There is no requirement that a sasCIF file would include all possible items defined in the dictionary. Similar to CIF, sasCIF is largely a free format.

4. Content of the sasCIF dictionary

As the sasCIF dictionary is an extension of the CIF dictionary, it is possible to use keywords defined in the CIF dictionary. Below only the items that need to be explicitly added to the sasCIF dictionary will be briefly described. For the description of the core CIF dictionary, see Hall S.R. *et al.*, (1991). In the following, the axial direction is defined to be perpendicular to the plane containing the incident beam and the scattered beam, so that the scattering vector is parallel to equatorial direction in the transmission geometry.

4.1. Beam geometry description.

`_beam.id` : The value of `_sas_beam.id` must uniquely identify the beam set-up used to collect each scattering data set. The SAS_DETC and SAS_SAMPLE categories point to this identifier. This structure permits e.g. to store the data from the same sample collected by X-ray and neutrons in a single file.

`_beam.monochromator_takeoff` : Twice the Bragg angle (2 theta) of the monochromator in degrees.

`_beam.radiation_pulse_duration` : Pulse duration in seconds.

`_beam.velocity_selector_speed` : Rotation speed of the velocity selector in rpm.

`_beam.velocity_selector_orientation` : Orientation of the velocity selector around the vertical axis in degrees.

`_beam.shape` : To allow accounting for other than rectangular beams; (i.e. circular) viewed at the sample.

`_beam.collimation_slit_size_ax`,
`_beam.collimation_slit_size_eq` : Defines the collimation slits in the axial and equatorial direction in millimetre.

`_beam.width_ax`, `_beam.width_eq` : Defines the beam size on the sample in the axial and equatorial directions in millimetre.

`_beam.divergence_ax`, `_beam.divergence_eq` : Defines the beam divergency in radians in the axial and equatorial planes.

`_beam.lambda_minimum` : Defines the minimum value of the wavelength λ for pulsed SANS.

`_beam.lambda_maximum` : Defines the maximum value of the wavelength λ for pulsed SANS.

`_beam.delta_lambda_over_lambda` : Defines the uncertainty (full width at half maximum $\delta\lambda / \lambda$) in the wavelength λ .

`_beam.dist_src/mono`, `_beam.dist_mono/spec`,
`_beam.dist_src/spec`, `_beam.dist_spec/anal`,
`_beam.dist_anal/detc`, `_beam.dist_spec/detc`

`_beam.dist_coll/spec` : Distances in meters for the instrument geometry from the radiation source to the monochromator, monochromator to the specimen, radiation source to the specimen, specimen to the analyser, analyser to the detector, specimen to the detector, and collimator to the specimen, respectively.

4.2. Detector description

`_detc.id` : The value of `_sas_detc.id` must uniquely identify the detector used to collect the given data set. A pointer in the SAS_SAMPLE category will point to this identifier. It permits to

store for instance the data from the same sample collected with a linear and a two-dimensional detector in the same file.

`_detc.beam_id` : This value is a pointer to the `_beam.id`.
`_detc.beamstop_shape` : Shape of the beamstop
`_detc.beamstop_size_ax`, `_detc.beamstop_size_eq`: Defines the beamstop size in the axial and equatorial directions in millimetre.
`_detc.beamstop_position_ax`,
`_detc.beamstop_position_eq`: Defines the beamstop position in the axial and equatorial directions in millimetre.
`_detc.pixnum_ax`, `_detc.pixnum_eq` : Defines the number of detector pixels in axial and equatorial direction.
`_detc.pixsize_ax`, `_detc.pixsize_eq` : Defines the pixel size in axial and equatorial direction in millimetre.
`_detc.center_ax`, `_detc.center_eq` : Defines the beam center on the detector in axial and equatorial direction in millimetre.
`_detc.radial_step` : Defines the radial step to group pixels for circular average in millimetre.
`_detc.merge_number` : Defines the number of channels merged together for the 1D-detector
`_detc.sector_width`, `_detc.sector_orientation` : Defines the sector width for sectorial average and the sector orientation with respect to equatorial direction in degrees.
`_detc.gamma_orientation`,
`_detc.beta_orientation`,
`_detc.alpha_orientation` : Define the eulerian angles of the orientation of the detector normal. When no rotation is applied, the normal to the detector is parallel to the primary beam direction.

4.3. Sample description

`_sample.id` : The value of `_sas_sample.id` uniquely identifies the sample name. A pointer in the SAS_INTENSITY category will point to this identifier. It permits to store for instance the data of a sample and of a matrix/buffer in the same file.
`_sample.beam_id` : The value is a pointer to `_beam.id` in the BEAM category.
`_sample.detc_id` : The value is a pointer to `_detc.id` in the DETC category.
`_sample.details` : A description of the sample such as the source of the sample, identification of standards, mixtures, etc.
`_sample.preparation_date` : The date of the sample preparation.
`_sample.specimen_concentration` : The specimen (e.g. protein in solution) concentration in mg/ml. The sample is composed by the specimen which is actually analysed and by the matrix/buffer.
`_sample.matrix_composition` : Description of the matrix or of the buffer surrounding the specimen.
`_sample.matrix_ph` : The pH of the matrix
`_sample.strain_description` : Description of the strain applied to the sample like uniaxial extension, uniaxial compression, fibers under dead-loading etc...
`_sample.strain_length_zero` : The length of the unstrained sample in millimetre.
`_sample.strain_length` : Assuming homogeneous deformation, the length of the compression or the extension applied to the sample in millimetre.
`_sample.measurement_date` : The date of the experiment.

`_sample.orientation` : Defines the angle between the sample normal and the incident beam in degrees.
`_sample.support` : A description of the holder where the sample is contained.
`_sample.thickness` : The thickness of the irradiated sample volume in millimetre.
`_sample.sample_transmission`,
`_sample.matrix_transmission`,
`_sample.background_transmission` : The transmission factors of the sample, of the matrix and of the background, respectively, defined as the intensity of transmitted beam divided by the intensity of incident beam.
`_sample.exposure_time` : The time during which the measurements were recorded in seconds
`_sample.calibration_details` : Description of the calibration used for normalisation of the scattering intensity (e.g. black carbon, water, standard protein).
`_sample.calibration_factor` : Multiplying the following data set by this value yields the intensity in absolute units.
`_sample.position_ax`, `_sample.position_eq`,
`_sample.position_z` : Defines the sample position in the three cartesian directions in millimetre.

4.4. Intensity description

`_intensity.sample_id` : This data item is a pointer to `.sample_id` in the sample category.
`_intensity.type` : Type of the data. The types allowed are "processed", "sample", "matrix", "background" and "detector response". If the type is "processed", the intensity I_p has already been processed.
The processing formula is $I_p = (I_s - I_m - I_b(T_m - T_s)) / (I_d * C * d_s * T_s)$ where I_s , I_m , I_b are the scattering intensity of the sample, the matrix and the background respectively. I_d is the detector response, T_s and T_m the transmission of the sample and of the matrix respectively, c is the concentration of the specimen, and D_s is the thickness of the sample. The types "sample", "matrix", "background" and "detector response" are the averaged raw data for the sample, matrix, background and detector response, respectively.
`_intensity.title` : data title
`_intensity.normalized_creator` : Name of the program used to process the data.
`_intensity.sample_raw` : Name of the file for the raw data of the sample.
`_intensity.matrix_raw` : Name of the file for the raw data of the matrix.
`_intensity.background_raw` : Name of the file for the raw data of the background.
`_intensity.sample_norm_file` : Name of the file for the normalised sample.
`_intensity.matrix_norm_file` : Name of the file for the normalised matrix.
`_intensity.background_norm_file` : Name of the file for the normalised background.
All the values of the following data items are usually written in a tabular form.
`_intensity.momentum_transfer` : Momentum transfer values $4\pi \sin(\theta)/\lambda$ in \AA^{-1}
`_intensity.intensity` : the scattering intensity.

The development of sasCIF dictionary is a community effort. The work was supported by the EU Grant B104-CT97-2143.

References

Bourne P. E., Berman H. M., McMahon B., Watenpaugh K. D., Westbrook J. & Fitzgerald P. M. D. (1997). *Methods in Enzymology*, **277**, 571-590

Hall S. R., Allen F. H. & Brown I. D. (1991). *Acta Cryst.* **A47**, 655-685

Hall S. R. (1993). *J. Appl. Cryst.*, **26**, 480.

Hall S. R. (1993). *J. Appl. Cryst.* **26**, 482.

Westbrook J.D. & Hall. S.R., (1995). Dictionary Description Language for Structure Macromolecular. Rutgers University, New Brunswick, NJ. Report NDB-110.

Westbrook J. D., H. Shu-Hsin & Fitzgerald P. M. D. (1997). *J. Appl. Cryst.* **30**, 79 - 83.