Argonne NATIONAL LABORATORY

The canSAS Format for Storage and Interchange of Reduced Multi-Dimensional Small-Angle Scattering Data

Presented by:

Pete R. Jemian

Advanced Photon Source, Argonne National Laboratory

(See final slide for complete author list.)

http://www.cansas.org

15th International Small-Angle Scattering Conference, http://sas2012.com

Session D05: Theory and Modelling 02, 2012-11-21, 1105 - 1120

Sydney Convention & Exhibition Centre, Sydney, Australia

18 - 23 November 2012

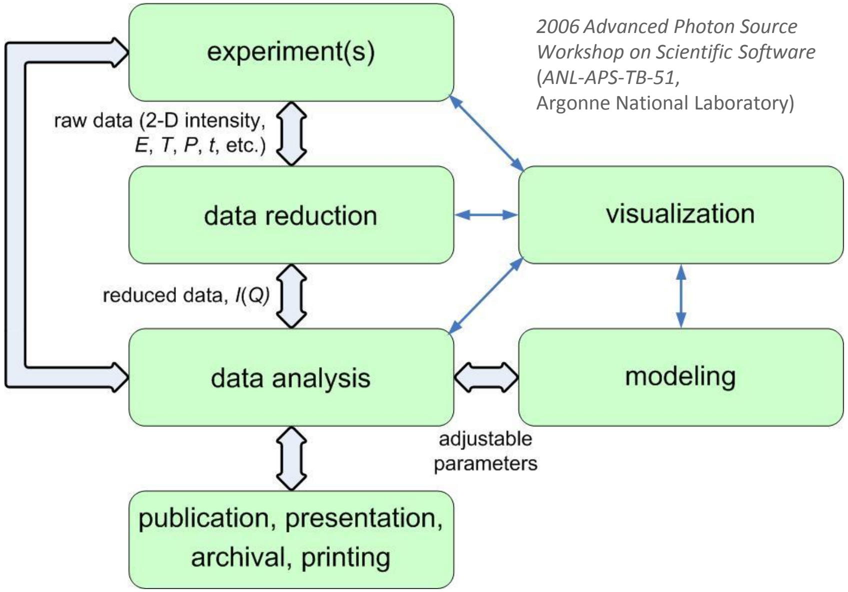
U.S. DEPARTMENT OF ENERGY

Scientific Benefits from these Efforts

- Improved interpretation methods and greater reliability of results
- Promote a greater understanding of the distribution and origin of uncertainty in data and metadata
- Establish a *defined interface* between the experiment and the analysis code
- Simplify the task for constructing data analysis software
- Facilitate routine analysis of large volumes of data
- Enable storage of appropriate metadata and uncertainties
- Meet standards for data deposition or publication
- More complete records of data provenance

- **Reduced data**, the target of the canSAS format, should be free from any correctable instrumental effects.
- The aim is for the canSAS format to be used in **Data Analysis** and **Data Deposition**.

Motivation



2006 Advanced Photon Source Workshop on Scientific Software (ANL-APS-TB-51, Argonne National Laboratory)

- One canSAS motive:
  - Provide better shared SAS data analysis software
- One means of doing so is through common data formats
  note: cansas1d/1.0, sasCIF

- For 2-D (and higher dimensionality), the job is harder
- Often, 2-D analysis software tries to start with raw data
- Data reduction steps are particular to the instrument *as it existed at one specific time*.

It is, and will always be, the responsibility of the instrument team to provide the process of converting the data measurements into **reduced data**.

Reduced data is the data presented for analysis after all instrument-specific artifacts and corrections have been applied.

The absolute minimum information required for the standard analysis of small-angle scattering measurements is intensity as a function of scattering vector, *I(Q)*.

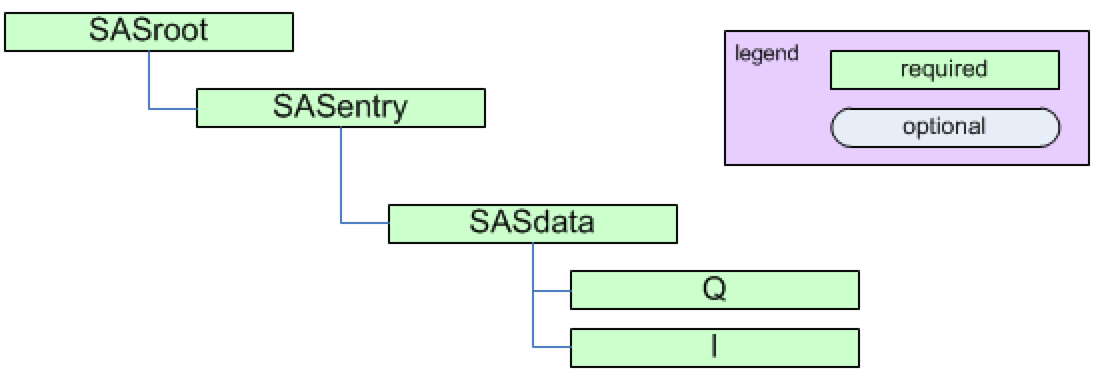
Requirements for the canSAS Format

- Allow for representation of *reduced data* of any dimensionality
  - 1D SAS data
  - 2D SAS data from detectors
  - additional dimensions for complex experiments or changing geometries
  - *Q* can be either a vector or a vector magnitude
- Identify and associate scanning axes
- Provide (when possible)
  - uncertainties and their constituents
  - masking information
- Allow for
  - complex experiments with multiple detectors
  - easy plotting of the data in close to their raw form
- Maintain the original dimensionality of the data if at all possible
- Use existing standards where possible or practical

Data Model

- General layout similar to canSAS1d/1.0
- Maps onto NeXus hierarchy directly
- Establishes *Q* and *I* as the absolute minimum content
- Adopt metadata from canSAS1d/1.0 (http://www.cansas.org/formats/canSAS1d/1.1/doc/overview.htm)
- Hierarchical structure (SASroot->SASentry->SASdata) allows adding data for further measurements, detectors, and interoperability with other formats

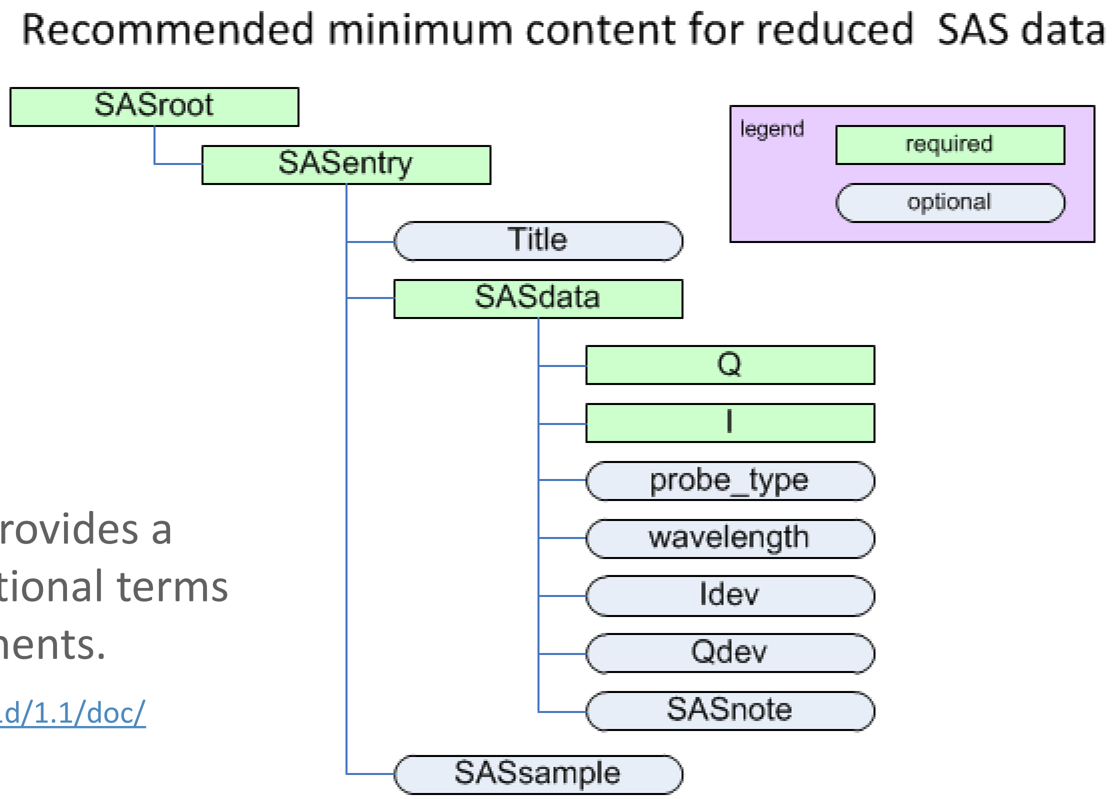
Absolute minimum requirement for analysis of SAS data



$I(Q)$

```
SASroot
  SASentry
    SASdata
      @axes=Q
      @Q_indices=0
      I: float[100]
      Q: float[100]
```

Additional Structure



The canSAS1d/1.0 format provides a dictionary of additional, optional terms for various types of experiments.

http://www.cansas.org/formats/canSAS1d/1.1/doc/

- Storing data from multiple detectors from the same sample and experiment is supported by either combining them into one dataset or providing multiple SASdata entries.

Multi-dimensional Data: A Simple Time-Series

- `@axes`
  - lists the axes of the `I` dataset (`Time` and `Q`)
  - associates the axes with the array indices
  - Only one index to use when looking up a `Q` value
- `@Q_indices` tells
  - lookup of *Q* depends on both the `Time` (0) and `Q` (1)
  - `Q` is time-dependent
- `Time` dataset provides the exposure timestamps
- Since there is no `Q` dataset, we find a *Q* vector
  - `Qx`, `Qy`, and `Qz` are provided
  - alternative would be a /*Q*/ term: `Q:float[4,35]`
- `I` provides the intensity array (reduced data)
- `Idev` provides the intensity uncertainties

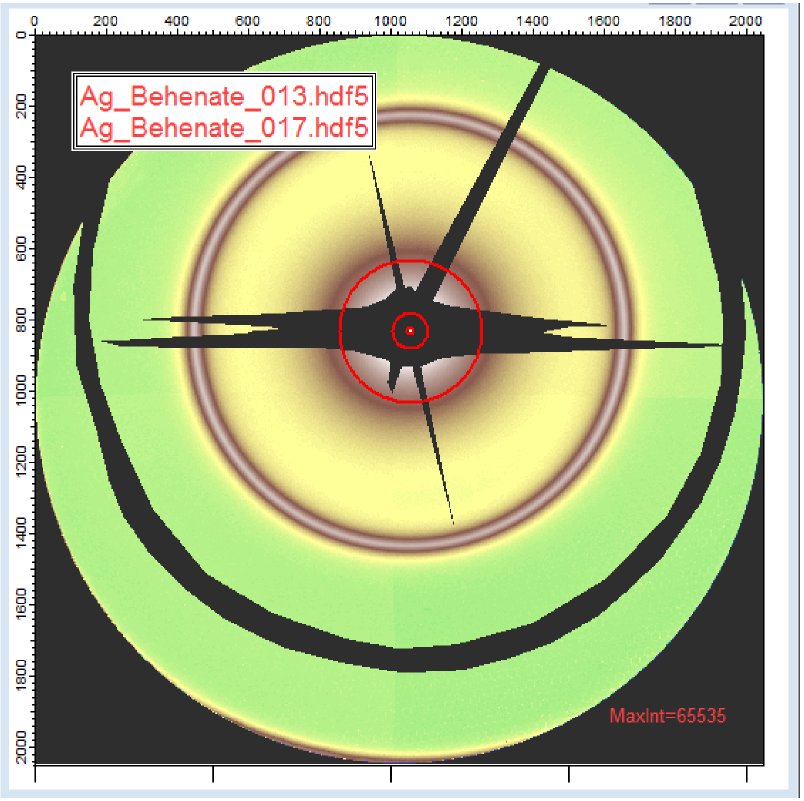
$$I(t,Q(t)) \pm \sigma(t,Q(t))$$

```
/ SASroot
  entry SASentry
    data SASdata
      @axes=Time,Q
      @Q_indices=0,1
      Qx: float[4,35]
      Qy: float[4,35]
      Qz: float[4,35]
      I: float[4,35]
        @uncertainty=Idev
      Idev: float[4,35]
      Time: float[4]
```

So for a given `i` and `j`, we find all the data:

`Qx[i,j], Qy[i,j], Qz[i,j], Time[i], I[i,j], Idev[i,j]`

Example with a data (detector) mask



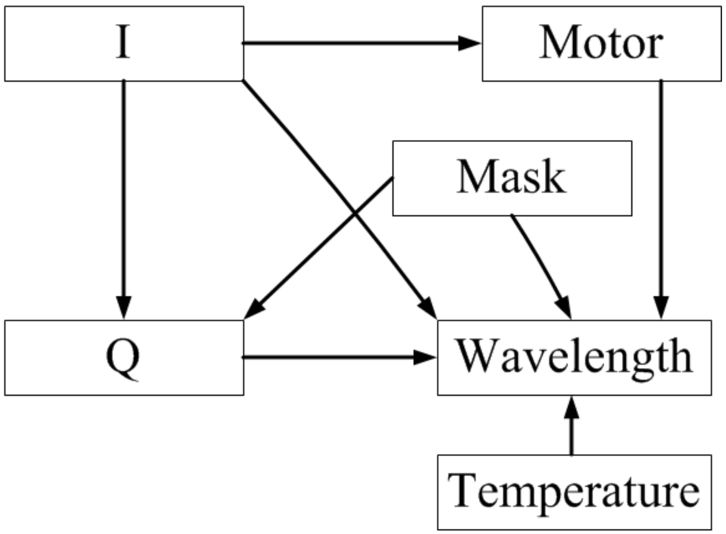
I(Q) and *mask(Q)*

```
/ SASroot
  entry SASentry
    data SASdata
      @axes=Q,Q
      @Q_indices=0,1
      @Mask_indices=0,1
      I: float[2048, 2048]
      Qx: float[2048, 2048]
      Qy: float[2048, 2048]
      Qz: float[2048, 2048]
      Mask: int[2048, 2048]
```

- Masking is important even for a reduced data format, especially when the original dimensionality is preserved. With multiple data dimensions, you may want to mask parts of the detector for only a subset of the exposures.
- This is a very simple example but it illustrates that the `Mask` is treated almost the same as any other axis.

Framework is Flexible

- 2-D Images sampled at different wavelengths and motor positions
- Temperature has been recorded for every exposure, so that this information is available for visualization and analysis in addition to the primary axes.
- Set of motor positions is different for each wavelength



I(m(w),w,Q(w)) and *Mask(q,Q(w), T(m,w)*

```
/ SASroot
  entry SASentry
    data SASdata
      @axes=Motor,Wavelength,Q,Q
      @Q_indices=1,2,3
      @Mask_indices=1,2,3
      @Motor_indices=0,1
      @Wavelength_indices=1
      @Temperature_indices=0,1
      I: float[m,w,128,512]
      Qx: float[w,128,512]
      Qy: float[w,128,512]
      Qz: float[w,128,512]
      Mask: float[w,128,512]
      Motor: float[m,w]
      Wavelength: float[w]
      Temperature: float[m,w]
```

Choice of File Storage Format

- Community is strongly divided between text files and binary files
- Both formats are very efficient for their purpose

- Requirements
  - Must be able to represent canSAS format as a structure
  - Must store primary data and metadata
  - Extraneous metadata should not be disruptive
  - Extensible (to store parameters and results of analyses)
  - Must have common support libraries

- Text files: XML (http://www.w3schools.com/xml)

- Binary files: HDF5 (http://www.hdfgroup.org/HDF5/)

- Other possibilities exist …

http://www.tumblr.com/tagged/turtles-all-the-way-down

Comments are Welcome!

- The canSAS format to store reduced data addresses the requirements adequately.
- The format is still in the phase for consultation and evaluation.
- More examples are available: http://www.cansas.org/formats/canSAS2012/1.0/doc/examples.html
- Comments are welcome.
- Also, an update to the 1-D XML format is just about ready: http://www.cansas.org/formats/canSAS1d/1.1/doc/

Recent work: 2012 canSAS workshop, Uppsala University, Uppsala, Sweden
http://www.cansas.org/wgwiki/index.php/canSAS-2012

Examples of the canSAS2012 data format

- $I(Q)$ models
  - 1-D $I(Q)$
  - 2-D image
  - 2-D SAS/WAS images
  - 2-D masked image
  - 2-D generic $I(Q)$
  - 2-D SANS and SAXS
  - several detectors
- $I(t,Q)$ models with time-dependence
  - 1-D $I(t,Q)$
  - 1-D $I(t,Q(t))$
  - 1-D $I(t,Q(t)) \pm \sigma(t,Q(t))$
  - 2-D $I(t,Q)$
  - 2-D $I(t,Q(t))$
  - 2-D $I(t,Q(t))$ masked image
- models with several varied parameters
  - 2-D $I(t,T,P,Q(t,T,P))$
  - 2-D $I(T,t,P,Q(t))$ images
- Unhandled Cases
  - 2-D image with Q_x & Q_y vectors

http://www.cansas.org

Thank You for your attention!
- from all the authors of this work

- P R **Jemian**, *Argonne National Laboratory, Advanced Photon Source, Argonne, IL 60439 USA*
- A J **Jackson**, *European Spallation Source ESS AB, PO Box 176, Lund 221 00, Sweden*
- S M **King**, *ISIS Facility, Science & Technology Facilities Council, Harwell Science & Innovation Campus, Didcot, Oxfordshire, OX11 0QX, UK*
- P R **Butler**, *National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899 USA*
- A R J **Nelson**, *Australian Nuclear Science and Technology Organisation, PMB1, Menai, Sydney, Australia*
- R E **Ghosh**, *Industrial Materials Group, Department of Chemistry, University College London, WC1H 0AJ, UK*
- T **Richter**, *Diamond Light Source Ltd., Diamond House, Harwell Science and Innovation Campus, Chilton, Didcot, Oxfordshire, OX11 0DE, U.K.*
- M **Doucet**, *Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA*
- A R **Rennie**, *Department of Physics and Astronomy, Uppsala University, 75120 Uppsala, Sweden*