NXcanSAS: Standard to Store Reduced SAS Data of any Dimension

A project of www.cansas.org Data Formats Working Group and www.nexusformat.org

Pete R. Jemian (APS), Andrew Jackson (ESS), Tobias Richter (ESS), Paul Butler (NIST), Steve King (ISIS), Adrian Rennie (Uppsala Univ.), and Brian Pauw (BAM),
APS: Advanced Photon Source, Argonne National Laboratory, Argonne, IL 60439, USA,

Goals

- Facilitate better sharing of SAS data analysis software
- Common data formats allow the easy use of different analysis software packages
- Generalize to describe simple experiments and complex experiments (such as with multiple detectors or multimodal experiments)
- Store reduced SAS data of any dimension
- Q can be either a vector (***Q***) or magnitude |**Q**|
- Identify and associate scanning axes (“self describing data”)
- Easy plotting of the data
- Maintain the original dimensionality of the data if at all possible
- Use existing standards where possible or practical
- Address the SAS community not reached by 2012 bioSAS standard
- Open source repositories

Reduced SAS Data

- Reduced SAS has (at minimum): *I(Q)*
- Data presented for analysis after all instrument-specific artifacts and corrections have been applied.
- The canSAS format is intended for use in Data Analysis and Data Deposition.

Scientific Benefits

- Establish a defined interface between experiment and analysis
- Enable storage of appropriate metadata and uncertainties
- Meet standards for data deposition or publication

Data format represents

- SAS data of any dimension
- Series:
  - Time
  - Temperature
  - Pressure
  - …
- uncertainties and resolution
- detector mask
- metadata:
  - Sample
  - Instrument
  - User
  - Facility
  - …
- Related data sets
- Multi-modal data (such as SAXS + WAXS)
- Analytical results

Decisions

- Use HDF5 (www.hdfgroup.org) …
- … with NeXus (www.nexusformat.org)

- Define ***NXcanSAS*** as NeXus application definition (http://tinyurl.com/z4thyrn) http://download.nexusformat.org/doc/html/classes/applications/NXcanSAS.html

Abstract

The communication of experimental results is common to scientific investigation. The method of presentation varies widely across investigation technique, and may stymie fundamental scientific goals such as sharing of results and replication of experiments.

Even within a limited community, such as small-angle scattering, the choice of how to organize stored information is fragmented, with the result that often, the data are deposited in ad hoc form. To increase access to data produced from publicly-funded research, funding agencies are now requiring that proposals describe how data will be made available.

With the increases in data volume due to higher-efficiency collection, increased experimental complexity, and larger and faster detectors, the plethora of ad hoc formats is a burden to the scientific community.

Reliance on a few, well-considered standards facilitates automated processes for analysis and correlation of scientific data. Furthermore, it leads to development of common tools for data visualization and analysis, and data catalogues for access, reference, and data mining.

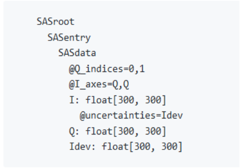
Other Information to be stored (when possible)

- additional dimensions for complex experiments (λ, T, t, P, …)
- uncertainties and their constituents
- masking information
- metadata (title, wavelength, radiation type and source, sample info, thickness, raw data reference, owner contact info…)
- analytical results
- complementary data

Schematic Examples

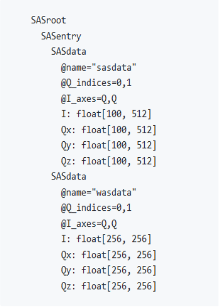
- Three brief schematics should illustrate how data is stored

2-D (image) I(|Q|) +/- sigma(|Q|)



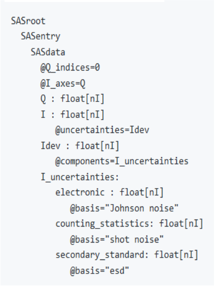
```
SASroot
  SASentry
    SASdata
      @Q_indices=0,1
      @I_axes=Q,Q
      I: float[300, 300]
        @uncertainties=Idev
      Q: float[300, 300]
      Idev: float[300, 300]
```

2-D SAS/WAS images



```
SASroot
  SASentry
    SASdata
      @name="sasdata"
      @Q_indices=0,1
      @I_axes=Q,Q
      I: float[100, 512]
      Qx: float[100, 512]
      Qy: float[100, 512]
      Qz: float[100, 512]
    SASdata
      @name="wasdata"
      @Q_indices=0,1
      @I_axes=Q,Q
      I: float[256, 256]
      Qx: float[256, 256]
      Qy: float[256, 256]
      Qz: float[256, 256]
```

Uncertainty Components



```
SASroot
  SASentry
    SASdata
      @Q_indices=0
      @I_axes=Q
      Q : float[nI]
      I : float[nI]
         @uncertainties=Idev
      Idev : float[nI]
         @components=I_uncertainties
      I_uncertainties:
         electronic : float[nI]
            @basis="Johnson noise"
         counting_statistics: float[nI]
            @basis="shot noise"
         secondary_standard: float[nI]
            @basis="esd"
```

Complementary work

- Worldwide Protein Data Bank:
  - Relies on a standardized data format
  - Large scale database of scientific data and metadata from a vast range of methods
  - available for research on the structures of proteins.
- bioSAS standard:
  - Guidelines for structural modelling of SAS from biomolecules in solution
  - adopted by the IUCr Commission on SAS in 2012

Background

- 1998: canSAS started on the idea of a common data format to aid the nomadic scatterer
- 2007: revisited this goal and agreed an XML format for 1D data
- 2009: canSAS1D/1.0 standard released
  - ISIS, NIST, ANSTO, APS, Diamond, ILL, Mantid, SASview, IgorPro (Irena & NIST macro sets), web-based converted for 3-column data, language interfaces (Fortran, Java, Python, …)
- 2012: agreed on standard for multi-dimensional data
- 2013: canSAS1D/1.1 standard released
- 2016: https://github.com/canSAS-org/NXcanSAS_examples
- 2017: NXcanSAS standard released

Compatible Software

- *Anything* that can read or write HDF5

- viewing: NeXpy, PyMCA
- reading: h5py, SASview, Mantid, IgorPro
- writing: h5py, SASview, Mantid

Status…

- v1.0 released 2017-01
- Provide example code to read and write
- Developers: obtain acceptance as storage format
- IUCr CSAS: obtain recognition as deposition format

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.