Australian Government

ANSTO

Standardising datasets

October 2019

Andrew Nelson

Australian Centre for Neutron Scattering

Science. Ingenuity. Sustainability.

Why?

Encouraging maximum use through:

- Inter-operability

  Data needs to use community agreed formats, language and vocabularies. The metadata will also need to use a community agreed standards and vocabularies, and contain links to related information using identifiers.

- Reusability

  Reusable data should maintain its initial richness. For example, it should not be diminished for the purpose of explaining the findings in one particular publication. It needs a clear machine readable licence and provenance information on how the data was formed. It should also have discipline-specific data and metadata standards to give it rich contextual information that will allow for reuse.

3.3.2.3. NXcanSAS

SAS community have gone through this process

Status:

application definition, extends NXobject

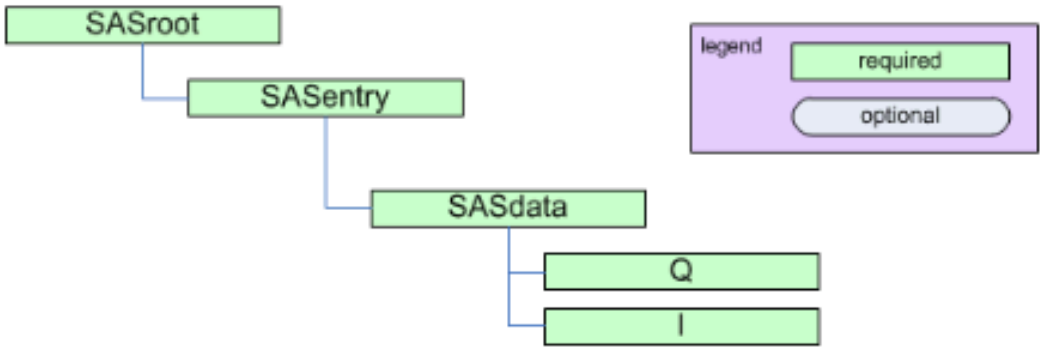
Description:

Implementation of the canSAS standard to store reduced small-angle scattering data of any dimension.

For more details, see:

- http://www.cansas.org/
- http://www.cansas.org/formats/canSAS1d/1.1/doc/
- http://cansas-org.github.io/canSAS2012/
- https://github.com/canSAS-org/NXcanSAS_examples

The minimum requirements for *reduced* small-angle scattering data as described by canSAS are summarized in the following figure:



The minimum requirements for *reduced* small-angle scattering data. (`full image`) See below for the minimum required information for a NeXus data file written to the NXcanSAS specification.

State of the art?

```
#RRF 1 0 Reflred 2003-11-05 for win
#date 2004-05-17
#title "may04 D2O TTAB#3 pH10"
#instrument NG7
#monitor 1.0
#temperature 0.000
#field 0.0000
#wavelength 4.760
#spec may04010.ng7 may04011.ng7
#back may04012.ng7
#slit may04005.ng7
# columns x y dy
0.005 0.986891613951 0.050345249368
0.0052 0.909864804934 0.040473037733
0.0054 0.976934317876 0.0452583055096
0.0056 0.800420953597 0.036123512147
0.0058 0.377786332409 0.0168653682439
0.006 0.242153622212 0.0119344902039
0.0062 0.152431605105 0.00674473783291
0.0064 0.124381806124 0.00589002614696
```

Pros

- What created me?
- Some sample details
- Temperature
- What files went into making me

Cons

- What are *x*/*y*/*dy*?
- No resolution information, *dx*
- Could the reduction be repeated from metadata?

ANSTO

State of the art?

6.035415258287789297e-03 1.016090004244090350e+00 1.868541816610796991e-02 2.427668326526606060e-04
6.146475778549943467e-03 1.019996753160879566e+00 1.648050700139233951e-02 2.473153619985038855e-04
6.259966760842364426e-03 1.021011632597536467e+00 1.505850094566428890e-02 2.519615368993668525e-04
6.375933082521700705e-03 1.021128901060202487e+00 1.421013036484228623e-02 2.567071782937978101e-04
6.494420599748085277e-03 1.052430246569295624e+00 1.401374195921436042e-02 2.615541463450624213e-04
6.615476164818657329e-03 1.063630540373261324e+00 1.359452685246213301e-02 2.665043411207850292e-04
6.739147646386259641e-03 1.049566384484101000e+00 1.293609810522867651e-02 2.715597033674042437e-04
6.865483947171630540e-03 1.023821488763809739e+00 1.210192009978146646e-02 2.767222152083363652e-04
6.994535023551020150e-03 1.068752675942207242e+00 1.193039640966239712e-02 2.819939009049613854e-04
7.126351905363263912e-03 1.027845241707705526e+00 1.060020952185197242e-02 2.873768276278781714e-04

Pros

- *x, y, dy* and *dx*

Cons

- No specified order of *x, y, dy, dx*
- No metadata
- What are *x, y, dy, dx*?
- What if dx isn't Gaussian?
- No way of repeating the reduction

ANSTO

State of the art?

```
<REFroot><REFentry time=”2013-01-22T11:01:56">
<Title>Ionic liquids</Title>
<REFsample>
<ID>silicon air</ID>
</REFsample>
<REFdata axes="Qz" rank="1" type="POINT" spin="UNPOLARISED" dim=”1">
<Run filename="[708, 709, 710]" preset="" size=”3”> </Run>
<R uncertainty="dR">1.0</R>
<Qz uncertainty="dQz" units="1/A">0.01</Qz>
<dR type="SD">0.00123</dR>
<dQz type="FWHM" units="1/A">0.0005</dQz>
</REFdata>
</REFentry></REFroot>
```

Pros

- *x, y, dy, dx*
- Units + some definition of dQz, dR
- Files that went into making it
- Can handle offspecular

Cons

- **No one else uses it**
- Complicated to read / write?

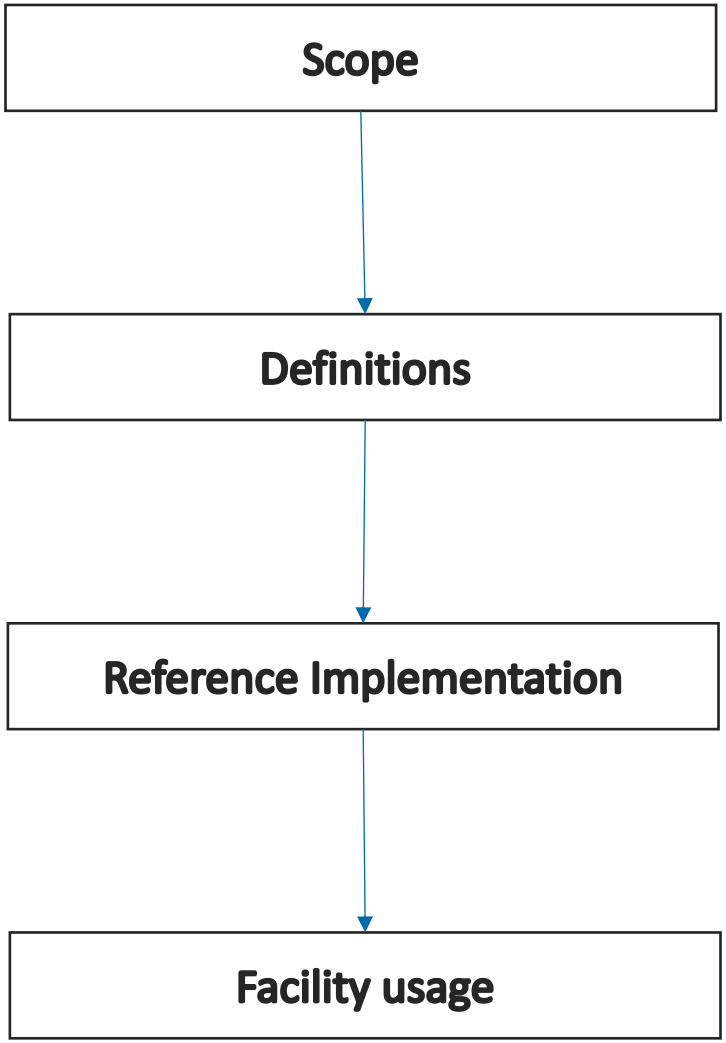
ANSTO

What do we want?

A dataset approach that :

- has an agreed set of definitions (what is *Q*, what is Q resolution?)
- is implemented by many facilities
- is used across a range of instruments
- is easily read by a variety of analysis packages

Roadmap



Note: no mention of file formats, definitions and scope are most important.
Implementation specifics come later

ANSTO

Discussion Topic 1 - Scope

For the use cases you're familiar with discuss the specific information of what needs to be in the dataset.
For example:

- Do we store Q or angles/wavelength?
- What level of instrument dependency remains?
  - are wavelength/angles required, or is Q/R ok? (c.f. footprint correction for scanning instruments)
- NR / XRR specifics
- Polarised Reflectometry
  - is polarization efficiency required?
  - S/P polarization?
- Resolution information
  - Probability distribution for each Q point?
- Required metadata
  - reduction program
  - what were the raw files?
  - sample footprint?
  - how to repeat the reduction?
  - background subtracted?
  - name/sample title?
  - temperature?
- Kinetic Data?
  - temperature/pressure/time/shear strain
- Offspecular?
- GISANS?
- Spin echo?
- RSoXRR?

Better to start out basic

ANSTO

Discussion Topic 2 – Common Definitions

For each piece of information that needs to be in the file write down its specific definition, down to its type.
Example:

We want to store momentum transfer, Q_z, perpendicular to the sample plane.
Defined as:

$$Q = \frac{2\pi}{\lambda}[\sin(2\theta - \Omega) + \sin\Omega]$$

Where Ω is the angle of incidence of the radiation (*degrees*), 2θ is the total deviation of the reflected beam in the scattering plane with respect to the incident beam (degrees), and λ is the wavelength of the radiation in *Angstrom*.
Units: 1 / Å
Storage: array of double, same shape as *R*

Only use integer/float/double, restricted set of labels. No complicated string parsing! No compound data types

ANSTO