Upgrade the standard text data file:

The canSAS standard for storing reduced one-dimensional small-angle scattering data in XML files

canSAS 1-D working group:

Pete R. Jemian (ANL), Andrew J. Jackson (NIST), Stephen M. King (ISIS), Ken C. Littrell (ORNL), Andrew R.J. Nelson (ANSTO), Ron E. Ghosh (ILL), Jan Ilavsky (ANL)

Presenter: Pete R. Jemian

XIV International Conference on Small-Angle Scattering
Sunday 13th - Friday 18th September, 2009
Oxford, UK

http://www.smallangles.net/wgwiki/index.php/cansas1d_documentation

Argonne
NATIONAL LABORATORY

... for a brighter future

U.S. Department of Energy

UChicago ►
Argonne LLC

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

canSAS 1DWG members

- Pete R. Jemian (ANL), Andrew J. Jackson (NIST), Stephen M. King (ISIS), Ken C. Littrell (ORNL), Andrew R.J. Nelson (ANSTO), Ron E. Ghosh (ILL), Jan Ilavsky (ANL)

Motivation

- Better sharing of SAS data analysis software

- Significant SAS community need satisfied by standardizing a format
  - robust
  - self-describing
  - text-based
  - reduced one-dimensional small-angle scattering data, *I(Q)*
  - Communicate data between users of our facilities

- XML
  - Sort-of easy to read
  - Strong support for languages
  - Strong support from community
  - Strong support for rules
  - Strong support for tools (yes, that rhymes)

Goal is to define such a format with:

- minimal verbosity
- leaves the data file instantly human-readable
- editable in simple editors
- importable by simple text import filters
- programs need not recognise
  - advanced structure in the file
  - require advanced programming interfaces.
- record of processing steps and analysis results

- The file should contain
  - primary data, *I(Q)*
  - metadata --- any other descriptive information about the sample, measurement, instrument, source, processing, or analysis steps.

- Compatible, where possible, with NeXus

Alternative data storage formats

- ASCII
  - Ubiquitous
  - Implied format, ill-defined
- NeXus: http://www.nexusformat.org
  - Data standard for neutron, electron, X-ray, muon science
  - Adopted by many radiation sources
  - PSI, Diamond, Soleil, just to name a few
  - HDF, defined for raw data, standardized for <u>instruments</u>
    (note that SAS is a technique, not an instrument)
- sasCIF: http://ww1.iucr.org/cif/spec
  - International Union of Crystallography standard
  - Intended as a 1D archive format for reduced data; ASCII, formatted
  - Used by ISIS/LOQ code (*COLETTE*) 2000-08 until 2007-summer
    (replaced by XML predecessor of this work)
  - flat file design, only 1 dataset/file

Justifications for XML

- Why XML?
  - data description language
  - not a file format
  - can be used to represent any data structure
  - Used by business
  - Extensible by design
  - Lots of tools available
  - Easy to translate using XSLT

- XML Schema Definition (XSD) language
  - strong data typing
  - syntax constraints
  - XSD written using XML
  - validate XML files

Anyone else using XML?

- **NeXus**
  - HDF (since beginning of NeXus)
  - XML (more recently)

- **GAML: Thermo Fisher Scientific**
  - *Bruker*
  - *Rigaku*

- **XRDML: Powder Diffraction (PANalytical)**

Argonne
NATIONAL LABORATORY

History & Timeline

- 2000: IUCr-CSAS established sasCIF, ISIS/LOQ adopted
- 2007-summer: SASXML installed at ISIS/LOQ
- 2007-10-31: canSAS workshop at NIST
- 2007-12-15: minimum XML data defined, start on XML Schema (XSD)
- 2008-12-21: pre-release version ready for committee to test
- 2008-02-26: various example datasets available on wiki
- 2008-03-07: project moved into public SVN (+ TRAC) repository
- 2008-03-14: IgorPro low-level import code available
- 2008-03-18: XSLT (web browser and Excel) available
- 2008-05-15: last change to cansas1d/1.0 standard (XSD)
- 2008-05-15: Java binding available
- 2008-05: IgorPro support in *Irena* available
- 2008-06-02: FORTRAN support
- 2008-08-04: adopted by ISIS/LOQ
- 2008-08-15: adopted by DANSE project (Python) for SANS analysis

Argonne
NATIONAL LABORATORY

Current Status

- **Adopted by facility:**
  - ISIS: default format for LOQ instrument, supported by *COLETTE*
  - NIST/NCNR:
  - DANSE:

- **Supported as an input format:**
  - Web browsers (through XSLT transformation to HTML or native XML)
  - Microsoft Excel
  - IgorPro *Irena* tool suite

- **Support has been implemented in these computer languages:**
  - IgorPro
  - Java (full XML binding available)
  - Python (part of DANSE and also separate binding through *GNOSIS*)
  - FORTRAN
  - PHP

The Minimum Set of information

- Easy to support legacy data

```
SASroot
  └ SASentry
      ├ Title
      ├ Run
      ├ SASdata
      │   └ Idata
      │       ├ Q
      │       └ I
      ├ SASsample
      │   └ ID
      ├ SASinstrument
      │   ├ name
      │   ├ SASsource
      │   │   └ radiation
      │   ├ SAScollimation
      │   └ SASdetector
      │       └ name
      └ SASnote
```
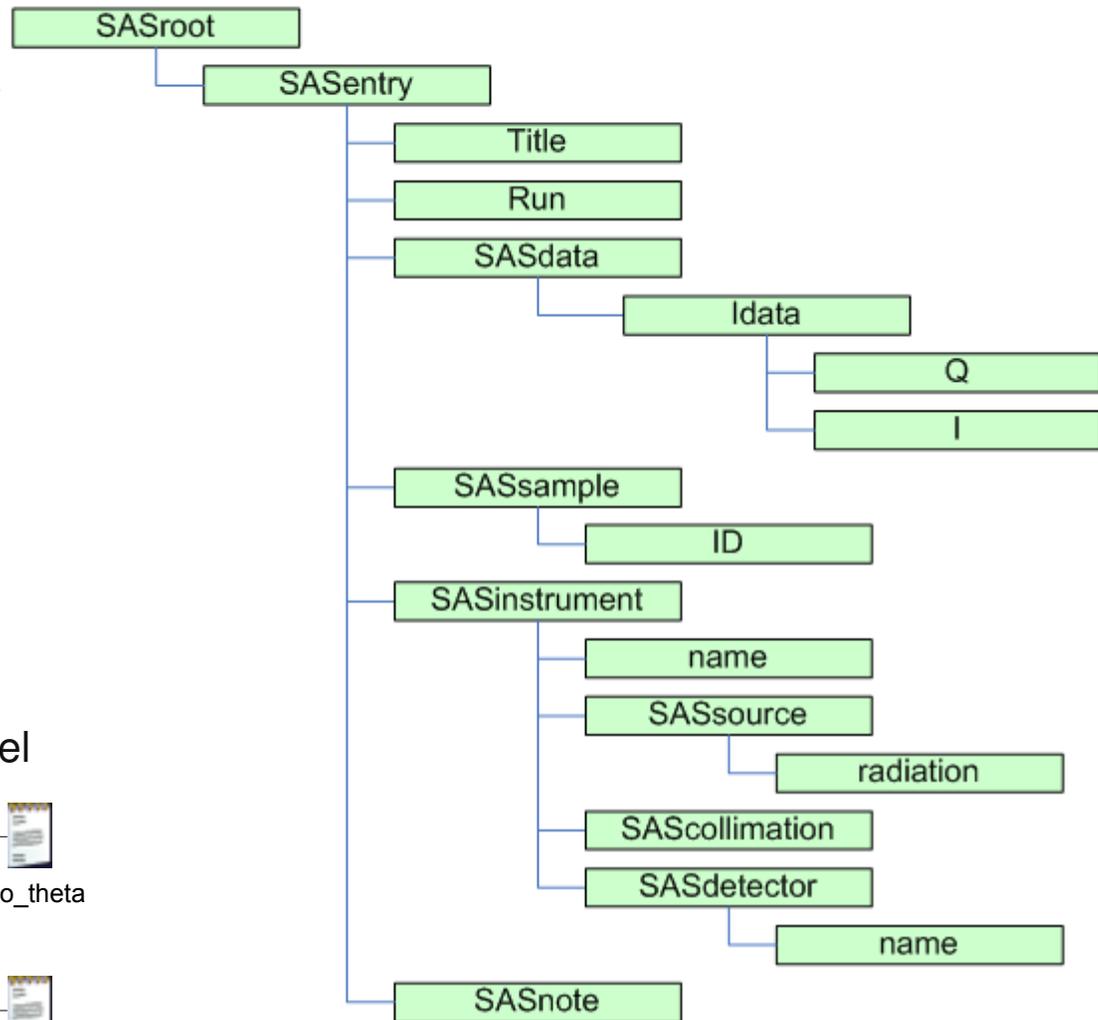
- Similar to basic NeXus model

```
NXroot ── NXentry ── NXdata ──┬── two_theta
                              └── counts
```

Argonne
NATIONAL LABORATORY
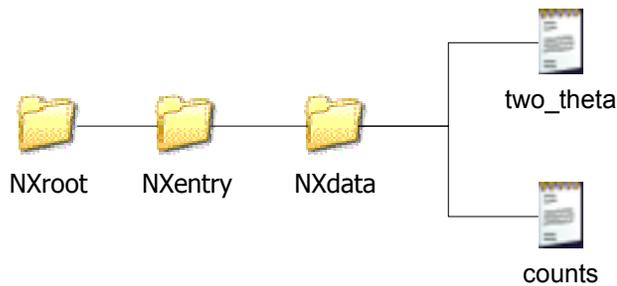
Where is the data? SASdata

Name	Type	occurrence	Description	Attributes
Q	floating-point number	[1..1]	$Q=(4\pi/\lambda)\sin(\theta)$ where λ is the wavelength of the radiation and 2θ is the angle through which the detected radiation has been scattered. The **unit** attribute is required. See cansas1d_documentation#Rules for acceptable values. Either *1/A* or *1/nm* are typical.	unit="{units}"
I	floating-point number	[1..1]	Intensity of the detected radiation. The **unit** attribute is required. See cansas1d_documentation#Rules for acceptable values. One possibility might be *1/cm*.	unit="{units}"
Idev	floating-point number	[0..1]	Estimated standard deviation of **I**. The **unit** attribute is required. See cansas1d_documentation#Rules for acceptable values. One possibility might be *1/cm*.	unit="{units}"

SASdata

Idata

Q

I

Idev

xor — Qdev / dQw / dQl

Qmean

Shadowfactor

<any>

<any>

Idata

legend
required
optional
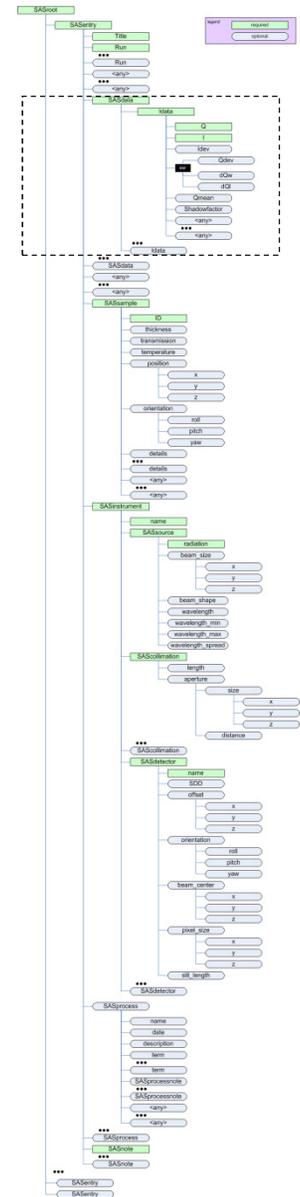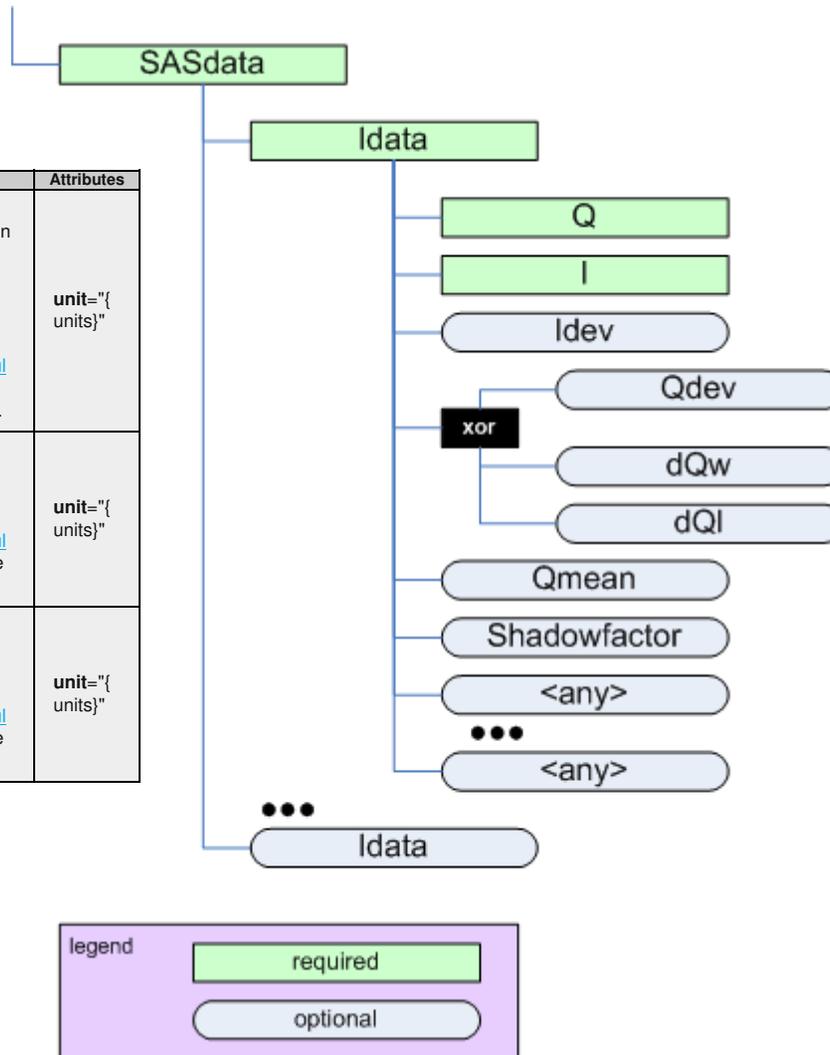
Example: 1994 SAXS data from dry chick collagen

- Raw data file

COLLAGEN.ASC	collagen.saxs

```
1    Sep 19 1994        01:41:02 am       Elt
2    ID: No spectrum identifier defined
3    Memory Size: 8192 Chls  Conversion
4
5    Chn      Counts  ROI
6       0,        0  000
7       1,        0  000
8       2,        0  000
9       3,        0  000
```

- ca.1994 reduced data file

COLLAGEN.ASC	collagen.saxs

```
1    dry chick collagen, d = 673 A
2    6531 eV, X6B
3
4    Q,1/□    SAXS      esd
5    0.022756    1107.6  8.5860
6    0.023296    1038.9  7.6445
7    0.023837    1071.0  7.9190
8    0.024377    1054.7  8.0684
9    0.024917    1061.3  8.2971
10   0.025457    1115.1  8.3305
11   0.025998    1276.1  8.5378
12   0.026538    1499.2  9.0048
13   0.027078    1738.2  10.172
14   0.027619    1802.5  10.335
15   0.028160    1738.5  10.130
```

canSAS XML data file: 1994 dry chick collagen SAXS


```
COLLAGEN.ASC   collagen.saxs   cs_collagen_full.xml

 1    <?xml version="1.0"?>
 2    <?xml-stylesheet type="text/xsl" href="example.xsl" ?>
 3    <SASroot version="1.0"
 4            xmlns="cansas1d/1.0"
 5            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 6            xsi:schemaLocation="cansas1d/1.0 http://svn.smallangles.net/svn/canSAS/1dwg/trunk/cansas1d.xsd"
 7            >
 8        <SASentry>
 9            <Title>dry chick collagen, d = 673 A, 6531 eV, X6B</Title>
10            <Run>Sep 19 1994      01:41:02 am</Run>
11            <SASdata>
12                <Idata><Q unit="1/A">0.022756</Q><I unit="a.u.">1107.6</I><Idev unit="a.u.">8.586</Idev><Qdev
13                <Idata><Q unit="1/A">0.023296</Q><I unit="a.u.">1038.9</I><Idev unit="a.u.">7.6445</Idev><Qde
14                <Idata><Q unit="1/A">0.023837</Q><I unit="a.u.">1071</I><Idev unit="a.u.">7.919</Idev><Qdev u
15                <!-- most data removed for clarity  -->
16            </SASdata>
17            <SASsample>
18                <ID>dry chick collagen, d = 673 A, 6531 eV, X6B</ID>
19            </SASsample>
20            <SASinstrument>
21                <name>X6B, NSLS, BNL</name>
22                <SASsource>
23                    <radiation>X-ray synchrotron</radiation>
24                    <wavelength unit="A">   1.898 <!-- = 12398/6531 --> </wavelength>
25                </SASsource>
26                <SAScollimation />
27                <SASdetector>
28                    <name>X6B PSD</name>
29                </SASdetector>
30            </SASinstrument>
31            <SASnote>
32                Sep 19 1994      01:41:02 am      Elt: 00090 Seconds
33                ID: No spectrum identifier defined
34                Memory Size: 8192 Chls  Conversion Gain: 1024  Adc Offset: 0000 Chls
35
36                dry chick collagen, d = 673 A
37                6531 eV, X6B
38            </SASnote>
39        </SASentry>
40    </SASroot>
41
```

Argonne NATIONAL LABORATORY

Various clients

Rules

- canSAS1d/1.0 XML data files will adhere to the standard if they can successfully validate against the established XML Schema (cansas1d.xsd)

- $Q = (4\pi / \lambda) \sin(\theta)$ where 2θ is scattering angle

- units to be given in standard SI abbreviations (*eg*, m, cm, mm, nm, K) with the following exceptions:
  - um (micrometres), C (celsius), A (Angstroms), percent (%), fraction, a.u. (arbitrary units), none (dimensionless, no relevant units)

- where reciprocal units need to be quoted, the format shall be "1/abbreviation"

- when raised to a power, use similar to "A^3" or "1/m^4" (and not "A3" or "m-4")

- x, y, & z axes: right-hand rule, origin not defined

- orientation (angles): roll, pitch, & yaw

- Unicode characters MUST NOT be used

- Binary data is not supported at this time

Just how hard can it be to get data into this format?
Fill out a simple WWW form
Leave blank what you don't know.

`http://www.smallangles.net/canSAS/xmlWriter`

- Enter basic SAS data (Q, I, and Idev) & some metadata
  - SAS data is likely copied from ASCII file in a text editor
  - Line breaks or commas are treated as white-space

canSAS1d/1.0 XML formatter

Title	This is a test
Run	not assigned
Sample ID	no sample
thickness (mm)	0
other sample details	DTE 27 mi
SAS instrument name	imaginary
"neutron" or "xray" (no quotes)	virtual
Wavelength (A)	1.54
Detector name	oldest
sample-detector distance (m)	0.3
SASnote	This shows how the XML formatter works
SAS data (3 columns: Q I Idev)	1 2 3, 4,5 6

◉ Styled ◯ Raw XML [Submit] [Reset]

Nicely formatted view
Save to samplename.xml from browser

SASentry1:(entry1)This is a test

SAS data			Selected Metadata		
(2 points)			**name**	**value**	**unit**
Q (1/A)	**I (1/cm)**	**Idev (1/cm)**	Title	This is a test	
1	2	3	Run	not assigned	
4	5	6	SASsample		
			SASsample_ID	no sample	
			SASsample_thickness	0	mm
			SASsample_details	DTE 27 mi	
			SASinstrument	imaginary	
			SASsource		
			SASsource_radiation	virtual	
			SASsource_wavelength	1.54	A
			SASdetector	oldest	
			SASdetector_SDD	0.3	m
			SASprocess	canSAS1d/1.0 XML formatter	
			SASprocess_date	2009-09-15 21:01:18	
			SASprocessnote	Id: index.php 87 2009-09-03 01:38:04Z prjemian	svnid
			SASprocessnote	formatting of text data into canSAS XML 1D standard	titleStr

Argonne
NATIONAL LABORATORY

Raw XML data

```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="cansasxml-html.xsl"?>
<SASroot xmlns="cansas1d/1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="1.0"
xsi:schemaLocation="cansas1d/1.0 http://svn.smallangles.net/svn/canSAS/1dwg/trunk/cansas1d.xsd">
<!--canSAS XML created by canSAS1d/1.0 XML formatter-->
  <SASentry name="entry1">
    <Title>This is a test</Title>
    <Run>not assigned</Run>
    <SASdata>
      <Idata> <Q unit="1/A">1</Q>    <I unit="1/cm">2</I>    <Idev unit="1/cm">3</Idev> </Idata>
      <Idata> <Q unit="1/A">4</Q>    <I unit="1/cm">5</I>    <Idev unit="1/cm">6</Idev> </Idata>
    </SASdata>
    <SASsample>
      <ID>no sample</ID>
      <thickness unit="mm">0</thickness>
      <details>DTE 27 mi</details>
    </SASsample>
    <SASinstrument>
      <name>imaginary</name>
      <SASsource>
        <radiation>virtual</radiation>
        <wavelength unit="A">1.54</wavelength>
      </SASsource>
      <SAScollimation/>
      <SASdetector>
        <name>oldest</name>
        <SDD unit="m">0.3</SDD>
      </SASdetector>
    </SASinstrument>
    <SASprocess>
      <name>canSAS1d/1.0 XML formatter</name>
      <date>2009-09-15 21:05:13</date>
      <SASprocessnote name="svnid">Id: index.php 87 2009-09-03 01:38:04Z prjemian</SASprocessnote>
      <SASprocessnote name="titleStr">formatting of text data into canSAS XML 1D standard</SASprocessnote>
    </SASprocess>
    <SASnote/>
  </SASentry>
</SASroot>
```

Argonne
NATIONAL LABORATORY

Higher dimensionality data

- 2-D area detectors
- Time series
- Other parametric studies

- Rely on NeXus
  - NeXus adopted by major facilities as standard (Diamond, Australian Light Source, others)
  - Waiting on next generation of NeXus
    - *NeXus is developing better way to describe "Application data"*
    - *Better support for how synchrotron X-ray data is collected*
    - *NXsas is one of the reasons for this change in the design of NeXus*
  - But NeXus is not quite ready yet so we are waiting …

Argonne
NATIONAL LABORATORY

Comments & Conclusions

- Good idea long time in the making
  - Agreement on standard between many user facilities
  - Flexible design tailored for specific community (small-angle)
  - Format allows access to data by variety of methods
  - Multiple measurements may be included within a single XML file
- Perceived competition with other standards bodies (NeXus, CIF)
  - Consider NeXus for 2-D (& higher) SAS data (see poster **PB58** today)
  - Possibly use NeXus/XML and base64binary encoding for big data?
- XML is a good method to store scientific data
  - Rich tools & support exist
  - XSLT to transform between different standards
- **The cansas1d/1.0 standard meets the objectives for a 1D standard, incorporating experiment metadata, and parameters and results of processing or analysis steps**.

http://www.smallangles.net/wgwiki/index.php/cansas1d_documentation

Thank you for your attention!

Argonne
NATIONAL LABORATORY