WP6: JRA on Provenance

Overview

Brian Matthews, Erica Yang

Scientific Applications Group
E-Science Centre
STFC Rutherford Appleton Laboratory

brian.matthews@stfc.ac.uk

WP 6 : JRA on provenance

- Start M7 (April 2012), Finish M30

– STFC (Lead)	18 SM
– ILL	6 SM
– ELETTRA	12 SM

WP6 : managing the data continuum

- The Provenance JRA
  - Extends the repository of information about an experiment
  - Tracking and logging the data analysis steps it links all the data artefacts
  - Records the data continuum
  - tracking of provenance of data
  - from proposal to publication.

- In general
  - A large and complex task
  - Establishing Science benefit

Objective

- To develop a conceptual framework, which can record and recall the data continuum, and especially the analysis process.
- To provide a software infrastructure which implements that model to record analysis steps hence enabling the tracing of the derivation of analysed data outputs.

Tasks

- **Task 1: Requirements for Provenance**
- **Task 2: Modelling the data continuum**
- **Task 3: Ontologies for specific instruments/techniques**
- **Task 4: Tool Support for the Data Continuum**
- **Task 5: Tracing the Data Continuum**
- **Task 6: Evaluation**

Deliverables

- Deliverables and month of delivery
- D6.1: Model of the data continuum in Photon and Neutron Facilities (M12)
- D6.2: Common ontology definition and definition of tools to support the use of provenance for Photon and Neutron Facilities (M18)
- D6.3: Tools for building research objects in Photon and Neutron Facilities (M24)
- D6.5: Evaluation report on provenance management in Photon and Neutron Facilities (M30) .

Requirements

- Explore some case studies e.g.
  - – ISIS – SNS (see later) (ISIS)
  - – Express services (ISIS, DLS?)
  - – DAWN (ESRF/DLS)
  - – Directly Programming Data Analysis Kit (DPDAK) (DESY)
  - – ISPyB (DLS/ESRF)
  - – Publication linking (e.g. DLS+IUCr, ISIS)
- Work with Virtual Laboratories

Data Continuum

ICAT
Metadata Repository

Experiment Proposal

ISIS

CCLRC ePublication Archive

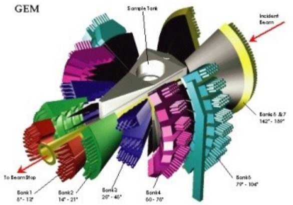
Welcome guest [sign in]

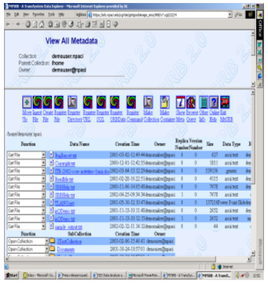
The CCLRC ePublication archive records the scientific output of CCLRC in the form of Journal Articles, Conference Papers, Technical Reports, ePrints, Theses and Books. To report any problems or send feedback please contact the ePublications Archive team at epublications@rl.ac.uk

Enter some text below to search the archive:

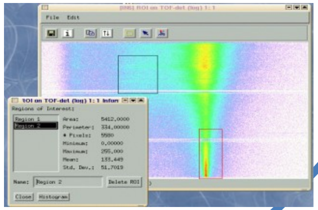
Or select one of the following browse indices:
by Year
by Author
by Journal
by Report Series
by Department
by Collaboration
by Type
by Title (with fulltext)

Create New Account





View All Metadata



Proposal

Approval

Scheduling

Experiment

Data cleansing

Data analysis

Record Publication

Scientist submits application for beamtime

Facility committee approves application

Facility registers, trains, and schedules scientist's visit

Scientists visits, facility run's experiment

Raw data filtered and cleansed

Tools for processing made available

Subsequent publication registered with facility

PanSoft

pandata europe

Science & Technology Facilities Council

Capturing data Provenance for Science: A Use Case and Next Steps

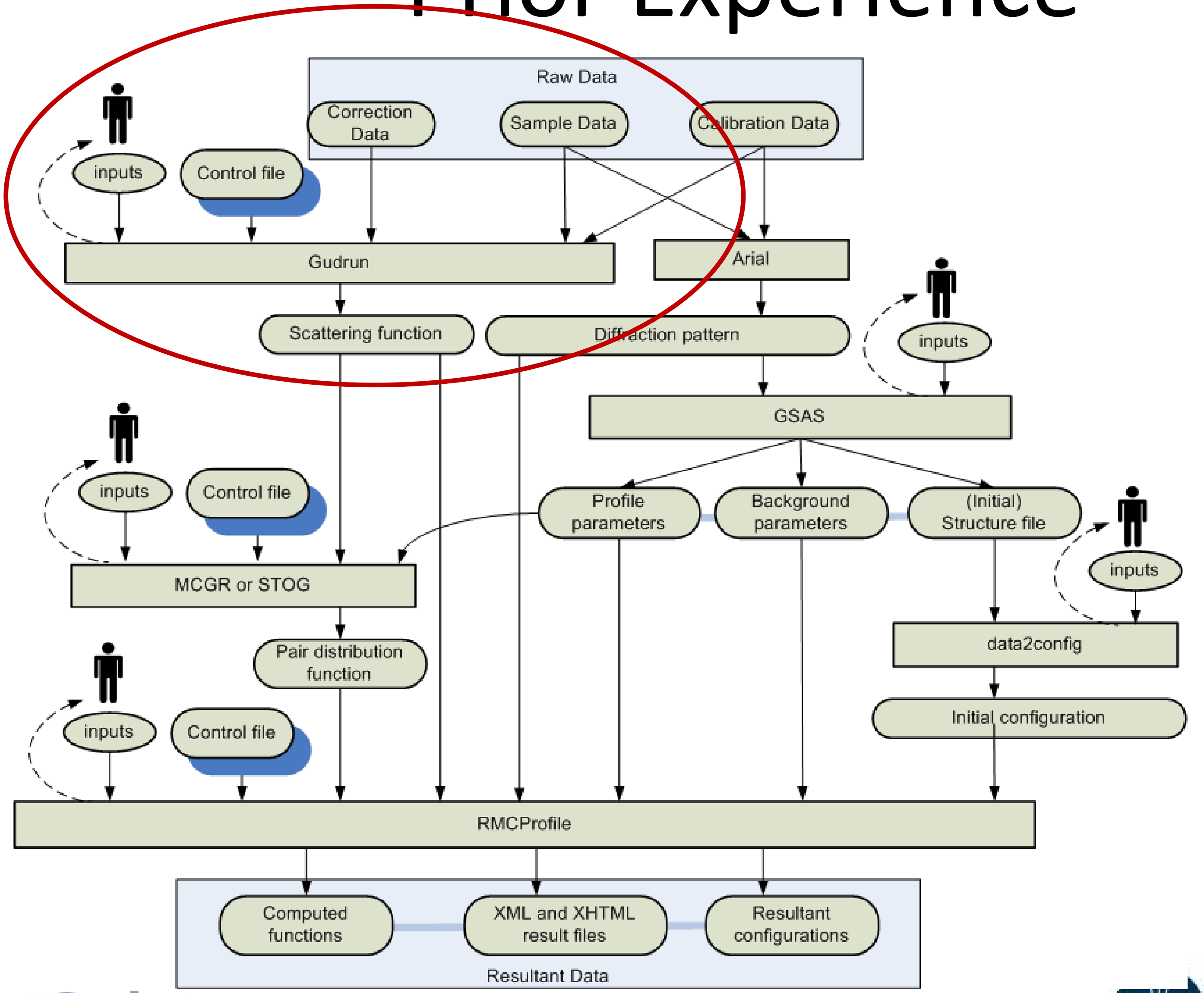
Erica Yang, Brian Matthews

Scientific Information Group
STFC e-Science

pandata europe

Science & Technology Facilities Council

Prior Experience



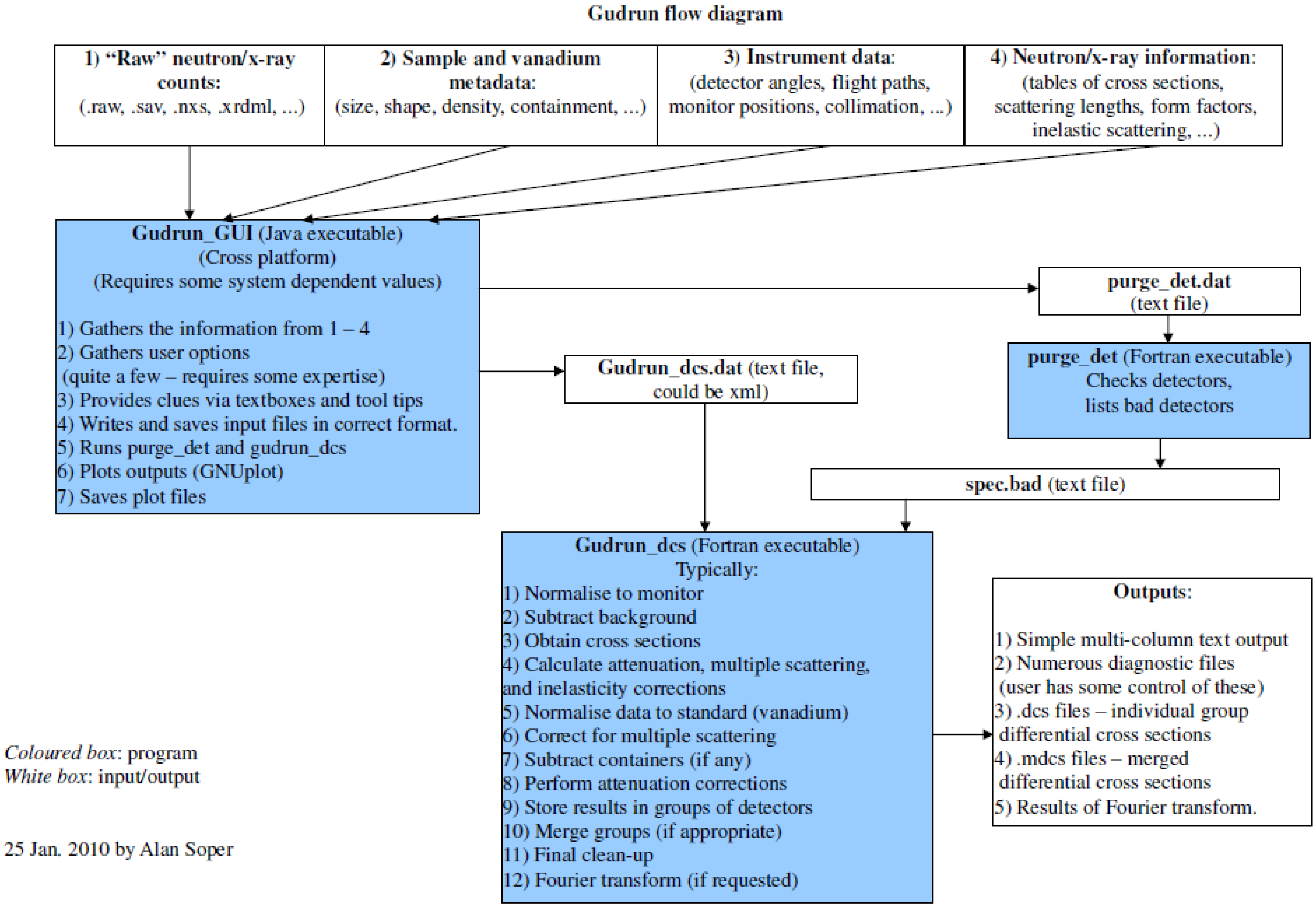
Raw data

Derived data

Resultant data

Credits: Martin Dove, Erica Yang (Nov. 2009)

Prior Experience



Gudrun flow diagram

1) "Raw" neutron/x-ray counts:
(.raw, .sav, .nxs, .xrdml, ...)

2) Sample and vanadium metadata:
(size, shape, density, containment, ...)

3) Instrument data:
(detector angles, flight paths, monitor positions, collimation, ...)

4) Neutron/x-ray information:
(tables of cross sections, scattering lengths, form factors, inelastic scattering, ...)

Gudrun_GUI (Java executable)
(Cross platform)
(Requires some system dependent values)

1) Gathers the information from 1 – 4
2) Gathers user options
(quite a few – requires some expertise)
3) Provides clues via textboxes and tool tips
4) Writes and saves input files in correct format.
5) Runs purge_det and gudrun_dcs
6) Plots outputs (GNUplot)
7) Saves plot files

purge_det.dat
(text file)

purge_det (Fortran executable)
Checks detectors,
lists bad detectors

Gudrun_dcs.dat (text file, could be xml)

spec.bad (text file)

Gudrun_dcs (Fortran executable)
Typically:
1) Normalise to monitor
2) Subtract background
3) Obtain cross sections
4) Calculate attenuation, multiple scattering, and inelasticity corrections
5) Normalise data to standard (vanadium)
6) Correct for multiple scattering
7) Subtract containers (if any)
8) Perform attenuation corrections
9) Store results in groups of detectors
10) Merge groups (if appropriate)
11) Final clean-up
12) Fourier transform (if requested)

Outputs:

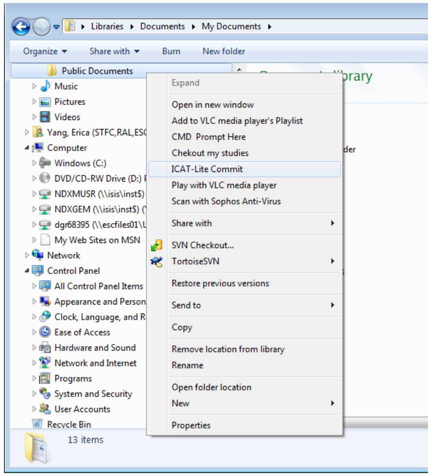
1) Simple multi-column text output
2) Numerous diagnostic files
(user has some control of these)
3) .dcs files – individual group differential cross sections
4) .mdcs files – merged differential cross sections
5) Results of Fourier transform.

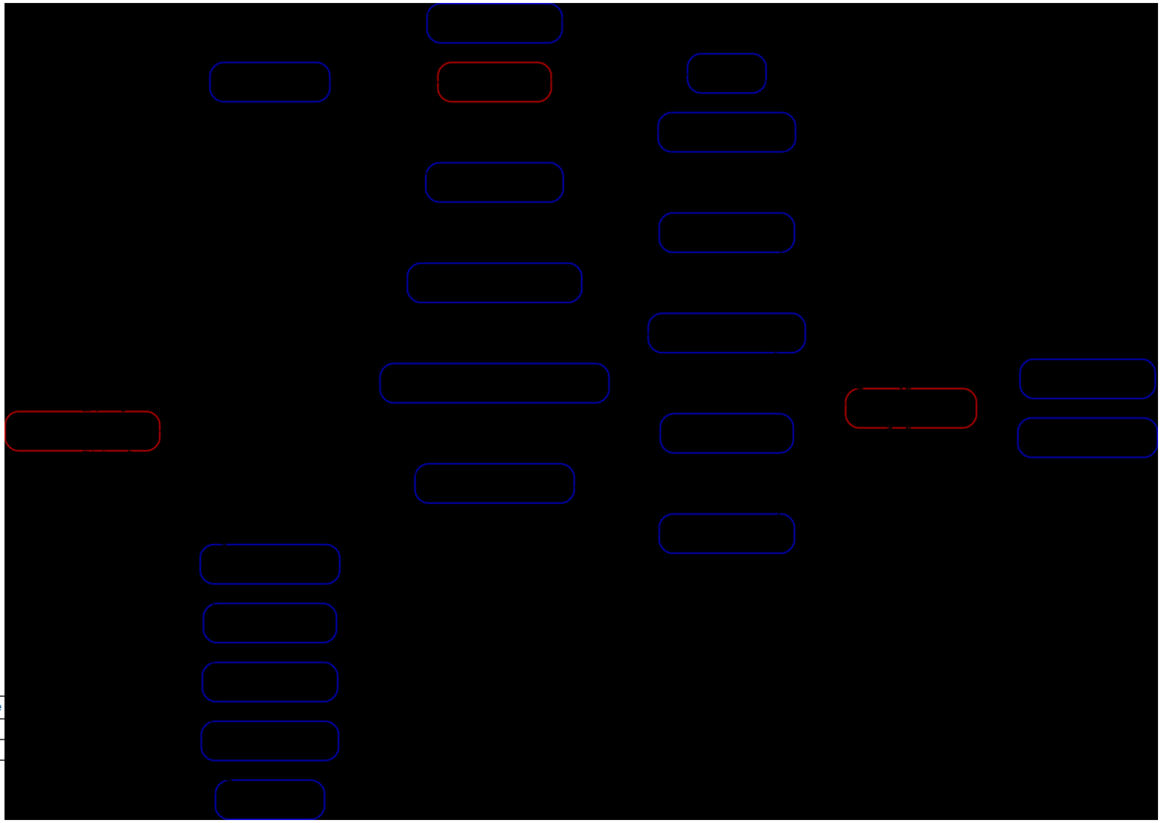
Coloured box: program
White box: input/output

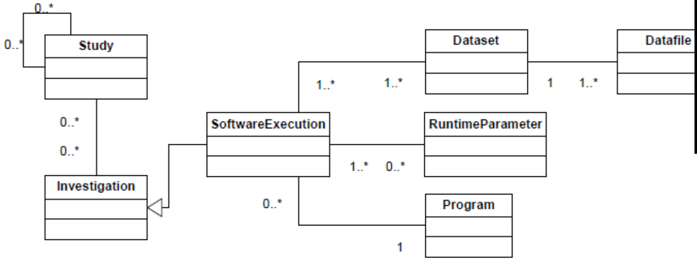
25 Jan. 2010 by Alan Soper

Credits: Alan Soper (Jan. 2010)

Prior Experience





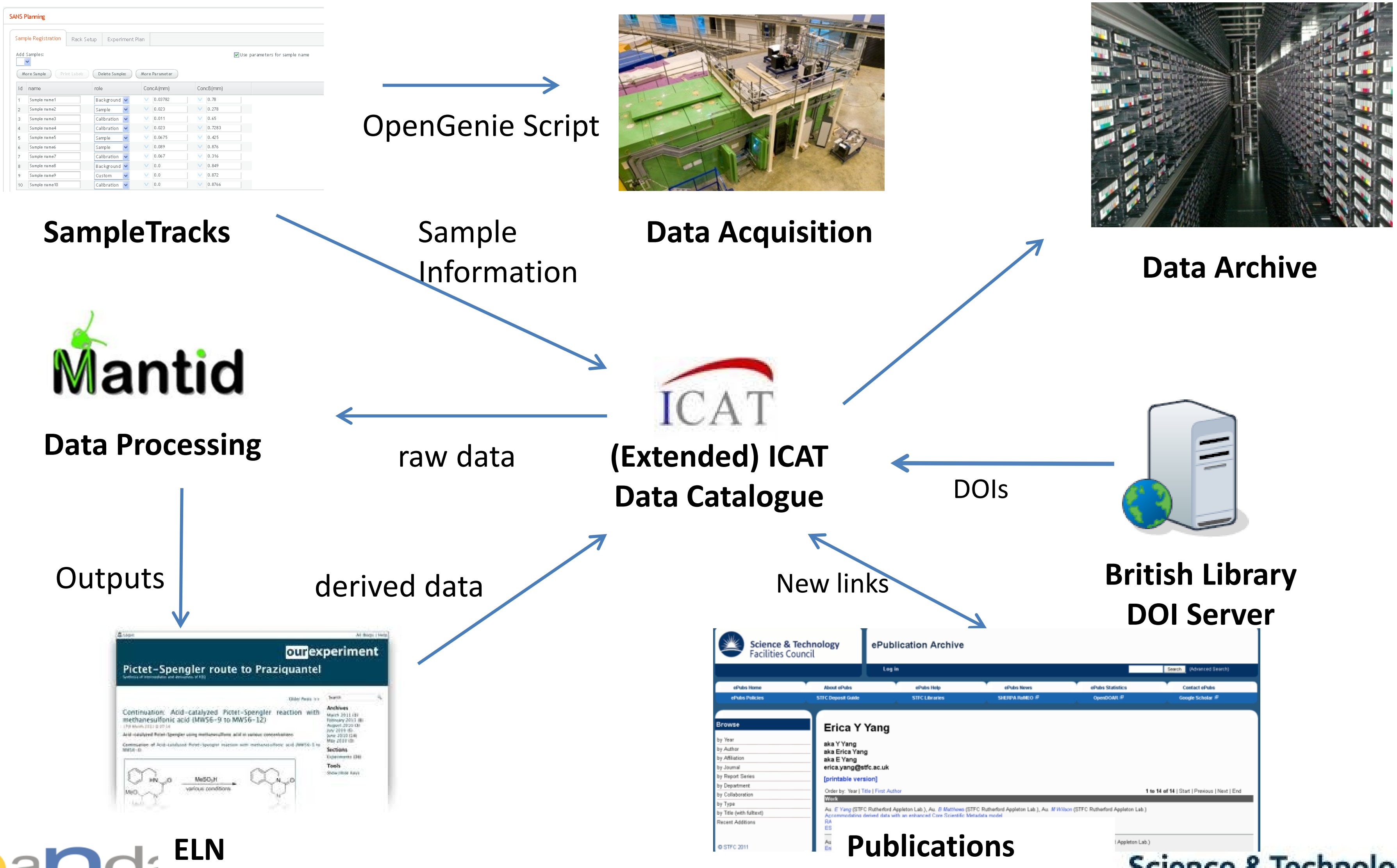


What we have learned

- Flexibility is the key to manage data provenance
  - Allow “Mix and match” of data processing trials
  - Support forward and backwards tracing of data provenance (e.g. raw<->derived<->paper)
- Researchers are hesitated to change their well established software/practice. “Why would I change?”
  - Need to demonstrate the benefits!

Use Case: Leveraging Data Provenance for Data Reduction

An Overview



SANS Planning

Sample Registration | Rack Setup | Experiment Plan

Add Samples:

☑ Use parameters for sample name

More Sample | Print Labels | Delete Samples | More Parameter

id	name	role	ConcA(mm)	ConcB(mm)
1	Sample name1	Background	0.03782	0.78
2	Sample name2	Sample	0.023	0.278
3	Sample name3	Calibration	0.011	0.65
4	Sample name4	Calibration	0.023	0.7283
5	Sample name5	Sample	0.0675	0.425
6	Sample name6	Sample	0.089	0.876
7	Sample name7	Calibration	0.067	0.316
8	Sample name8	Background	0.0	0.849
9	Sample name9	Custom	0.0	0.872
10	Sample name10	Calibration	0.0	0.8766

ourexperiment

Pictet-Spengler route to Praziquantel

Continuation: Acid-catalyzed Pictet-Spengler reaction with methanesulfonic acid (MWS6-9 to MWS6-12)

Archives

Sections

Tools

Science & Technology Facilities Council

ePublication Archive

Log in | Search | (Advanced Search)

ePubs Home | About ePubs | ePubs Help | ePubs News | ePubs Statistics | Contact ePubs

ePubs Policies | STFC Deposit Guide | STFC Libraries | SHERPA RoMEO | OpenDOAR | Google Scholar

Browse

by Year
by Author
by Affiliation
by Journal
by Report Series
by Department
by Collaboration
by Type
by Title (with fulltext)
Recent Additions

© STFC 2011

Erica Y Yang

aka Y Yang
aka Erica Yang
aka E Yang
erica.yang@stfc.ac.uk

[printable version]

Order by: Year | Title | First Author — 1 to 14 of 14 | Start | Previous | Next | End

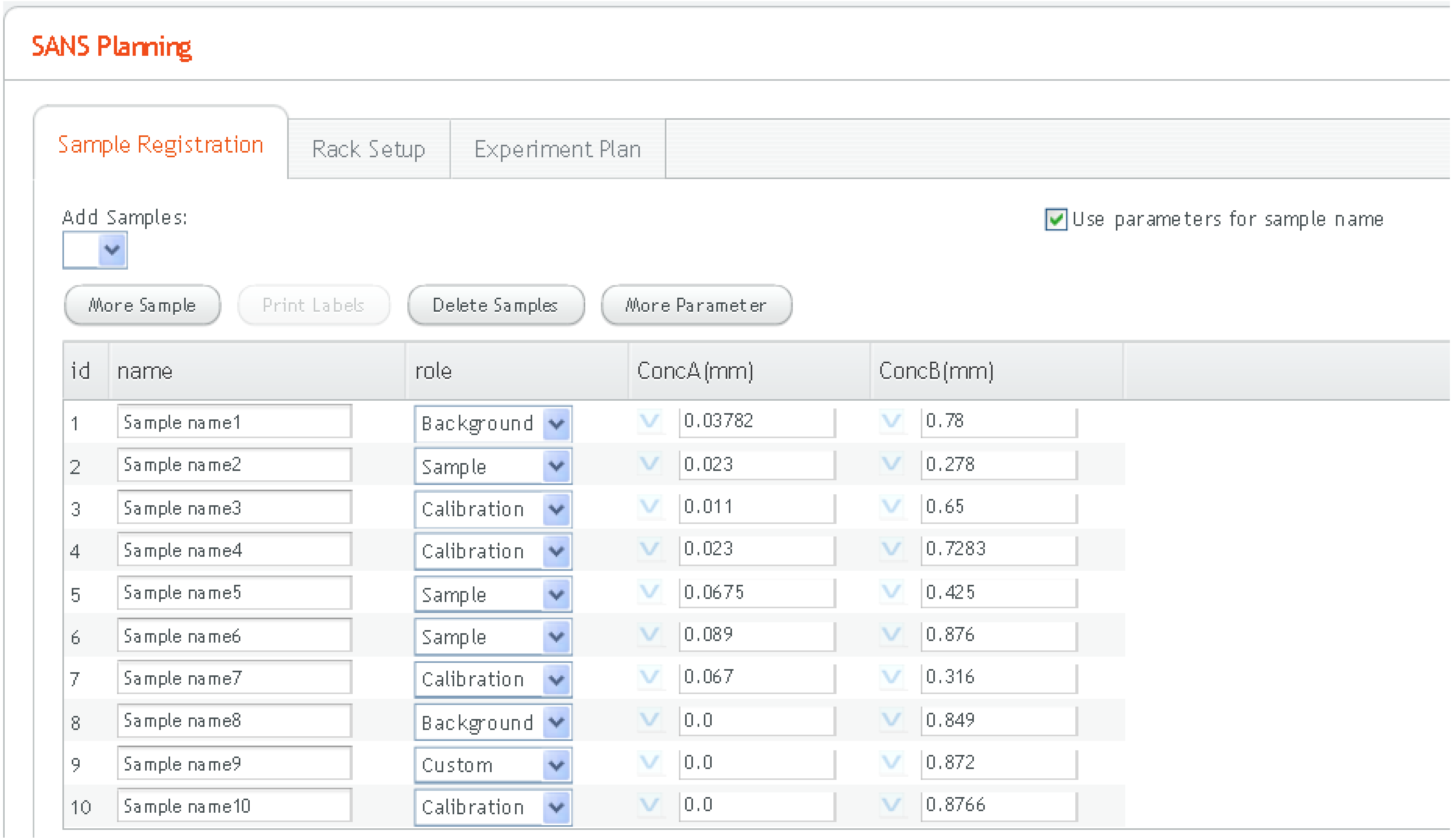
Work

Au. E Yang (STFC Rutherford Appleton Lab.), Au. B Matthews (STFC Rutherford Appleton Lab.), Au. M Wilson (STFC Rutherford Appleton Lab.)
Accommodating derived data with an enhanced Core Scientific Metadata model

pandata europe

Science & Technology Facilities Council

Sample Registration



Add Samples:

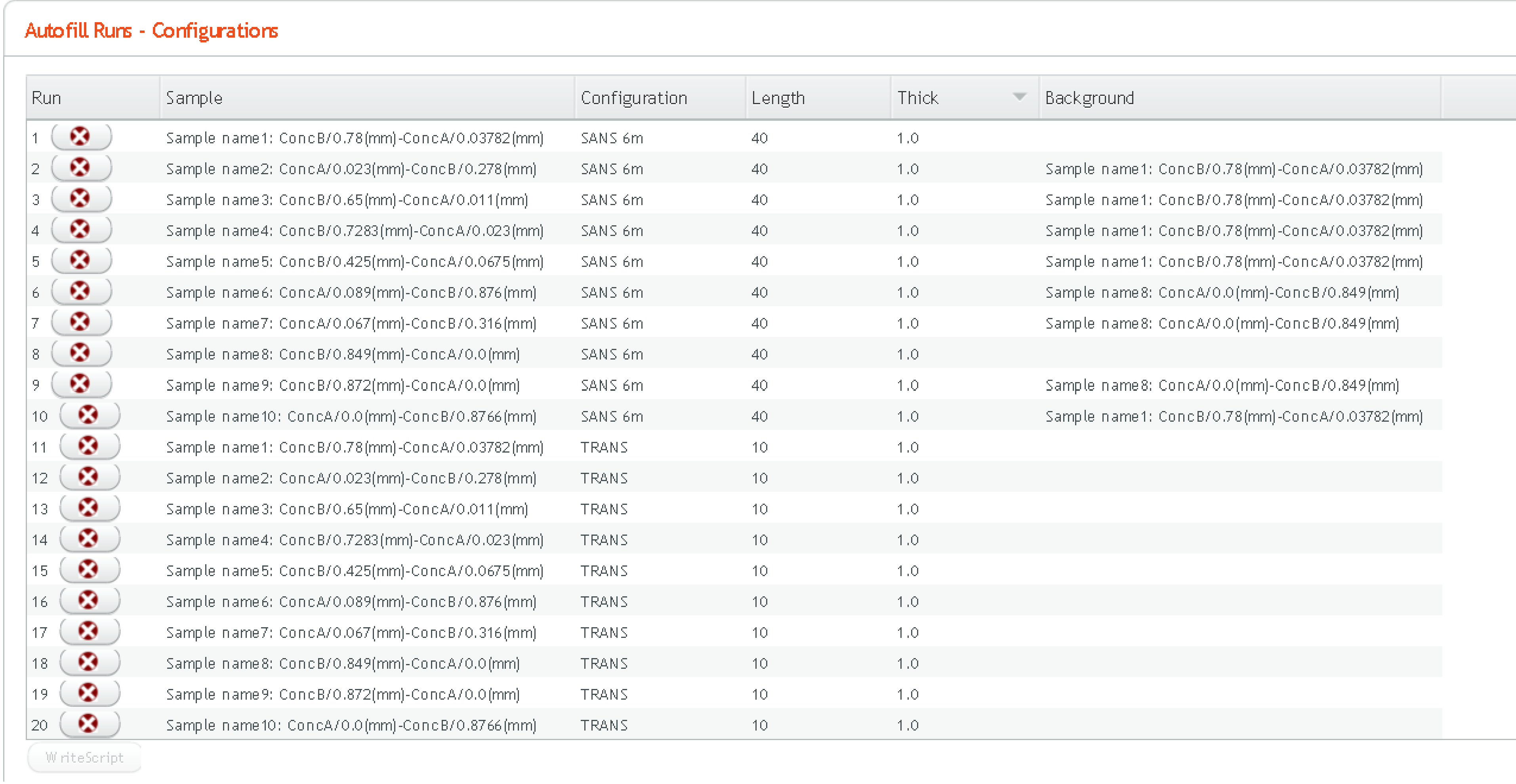
☑ Use parameters for sample name

id	name	role	ConcA (mm)	ConcB(mm)
1	Sample name1	Background	0.03782	0.78
2	Sample name2	Sample	0.023	0.278
3	Sample name3	Calibration	0.011	0.65
4	Sample name4	Calibration	0.023	0.7283
5	Sample name5	Sample	0.0675	0.425
6	Sample name6	Sample	0.089	0.876
7	Sample name7	Calibration	0.067	0.316
8	Sample name8	Background	0.0	0.849
9	Sample name9	Custom	0.0	0.872
10	Sample name10	Calibration	0.0	0.8766

pandata europe

Science & Technology Facilities Council

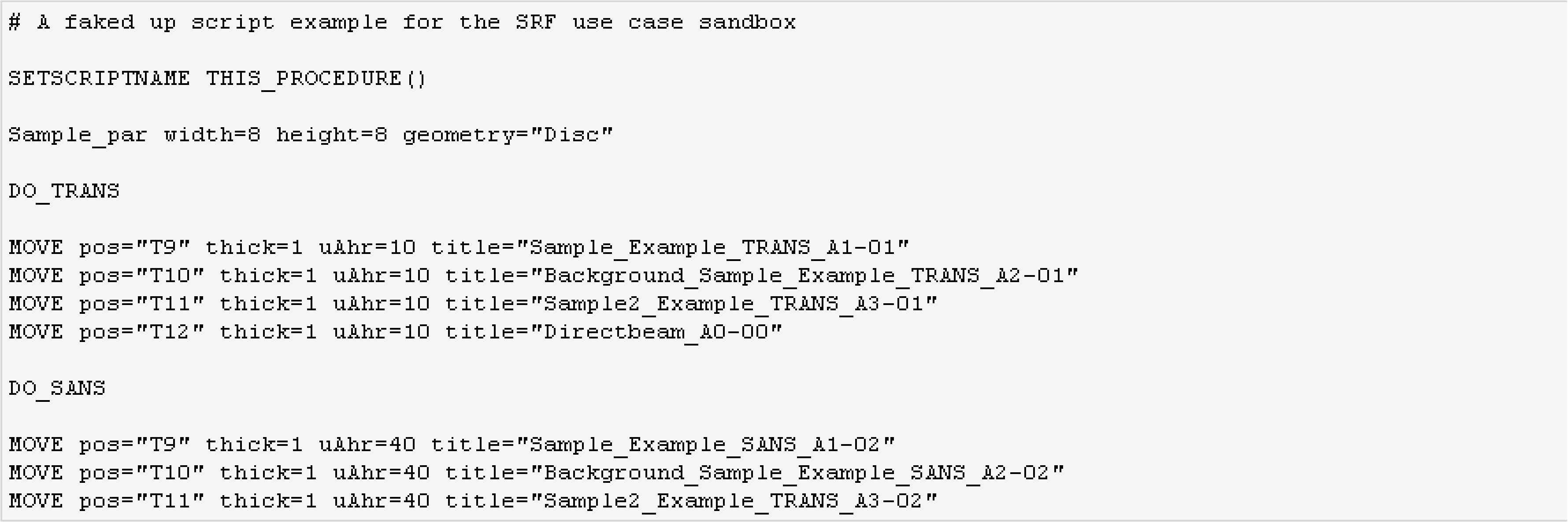
Experiment Planning – Runs: Detailed View



Autofill Runs - Configurations

Run	Sample	Configuration	Length	Thick	Background
1 ⊗	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)	SANS 6m	40	1.0	
2 ⊗	Sample name2: ConcA/0.023(mm)-ConcB/0.278(mm)	SANS 6m	40	1.0	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)
3 ⊗	Sample name3: ConcB/0.65(mm)-ConcA/0.011(mm)	SANS 6m	40	1.0	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)
4 ⊗	Sample name4: ConcB/0.7283(mm)-ConcA/0.023(mm)	SANS 6m	40	1.0	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)
5 ⊗	Sample name5: ConcB/0.425(mm)-ConcA/0.0675(mm)	SANS 6m	40	1.0	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)
6 ⊗	Sample name6: ConcA/0.089(mm)-ConcB/0.876(mm)	SANS 6m	40	1.0	Sample name8: ConcA/0.0(mm)-ConcB/0.849(mm)
7 ⊗	Sample name7: ConcA/0.067(mm)-ConcB/0.316(mm)	SANS 6m	40	1.0	Sample name8: ConcA/0.0(mm)-ConcB/0.849(mm)
8 ⊗	Sample name8: ConcB/0.849(mm)-ConcA/0.0(mm)	SANS 6m	40	1.0	
9 ⊗	Sample name9: ConcB/0.872(mm)-ConcA/0.0(mm)	SANS 6m	40	1.0	Sample name8: ConcA/0.0(mm)-ConcB/0.849(mm)
10 ⊗	Sample name10: ConcA/0.0(mm)-ConcB/0.8766(mm)	SANS 6m	40	1.0	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)
11 ⊗	Sample name1: ConcB/0.78(mm)-ConcA/0.03782(mm)	TRANS	10	1.0	
12 ⊗	Sample name2: ConcA/0.023(mm)-ConcB/0.278(mm)	TRANS	10	1.0	
13 ⊗	Sample name3: ConcB/0.65(mm)-ConcA/0.011(mm)	TRANS	10	1.0	
14 ⊗	Sample name4: ConcB/0.7283(mm)-ConcA/0.023(mm)	TRANS	10	1.0	
15 ⊗	Sample name5: ConcB/0.425(mm)-ConcA/0.0675(mm)	TRANS	10	1.0	
16 ⊗	Sample name6: ConcA/0.089(mm)-ConcB/0.876(mm)	TRANS	10	1.0	
17 ⊗	Sample name7: ConcA/0.067(mm)-ConcB/0.316(mm)	TRANS	10	1.0	
18 ⊗	Sample name8: ConcB/0.849(mm)-ConcA/0.0(mm)	TRANS	10	1.0	
19 ⊗	Sample name9: ConcB/0.872(mm)-ConcA/0.0(mm)	TRANS	10	1.0	
20 ⊗	Sample name10: ConcA/0.0(mm)-ConcB/0.8766(mm)	TRANS	10	1.0	

Instrument Control Script



```
# A faked up script example for the SRF use case sandbox

SETSCRIPTNAME THIS_PROCEDURE()

Sample_par width=8 height=8 geometry="Disc"

DO_TRANS

MOVE pos="T9" thick=1 uAhr=10 title="Sample_Example_TRANS_A1-01"
MOVE pos="T10" thick=1 uAhr=10 title="Background_Sample_Example_TRANS_A2-01"
MOVE pos="T11" thick=1 uAhr=10 title="Sample2_Example_TRANS_A3-01"
MOVE pos="T12" thick=1 uAhr=10 title="Directbeam_A0-00"

DO_SANS

MOVE pos="T9" thick=1 uAhr=40 title="Sample_Example_SANS_A1-02"
MOVE pos="T10" thick=1 uAhr=40 title="Background_Sample_Example_SANS_A2-02"
MOVE pos="T11" thick=1 uAhr=40 title="Sample2_Example_TRANS_A3-02"
```

Flexible Data Model

- Arbitrary number and types of sample parameters, for different types of experiments
- Arbitrary number and types of sample configurations, for different types of instruments
- Designed to work with existing instrument control system, data acquisition system, data catalogue (ICAT), and data processing software
- Designed to allow capturing and tracking of data provenance from samples, to raw and reduced data, and to publications, and backwards

Next Steps

- Firming up the use cases
  - Revisiting the requirements and existing approaches for data provenance within the PanData facilities
  - Modelling the data continuum
    - Commonalities across existing approaches
    - Common processes across facilities
  - Work with other WPs/partners, e.g.
    - WP4: Data Catalogue Service
    - WP5: Use cases in Virtual Laboratories
    - Directly Programming Data Analysis Kit (DPDAK)
    - Nexus Application Definitions
  - Flagship demonstrations to demonstrate the benefits
    - Express services

Benefits

- Showcase that data provenance can directly improve research productivity
- Improve facility operational efficiency
- Follow the “non-intrusive” principle to capture and catalogue metadata, designing as part of researchers’ existing workflow