Australian Government | ANSTO

Scientific Computing and Data Management at the Australian Synchrotron

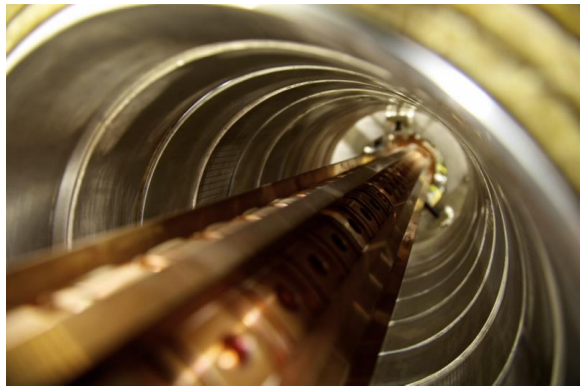
IUCr 2023 – Data Workshop, 22/08/2023

Dr Andreas Moll

Manager – Scientific Computing

Science. Ingenuity. Sustainability.

ANSTO

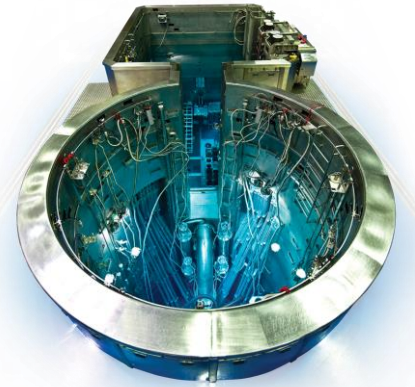


Clayton | VIC

Australian Synchrotron

Lucas Heights | NSW

Main campus



OPAL
multi-purpose reactor

ANSTO

Australian Synchrotron

A research facility

- Particle accelerator with 216m circumference
- Beam available 24 hours, 6 days a week

A user focused facility

- 5500+ visits per year
- 10 (+2) operating experimental end stations (beamlines)
- 586 Journal Publications in 2022
- Generate 1.5 PB of data each year

ANSTO

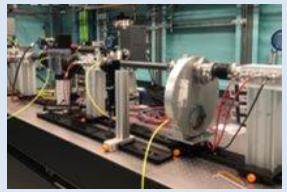
Australian Synchrotron & BRIGHT

A growing facility (BRIGHT program)

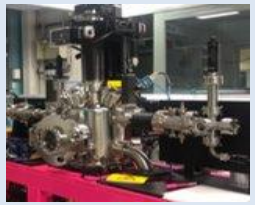
- 10 original beamlines
- 8 new beamlines
- 3 already operating

- Opportunity to “refresh” software

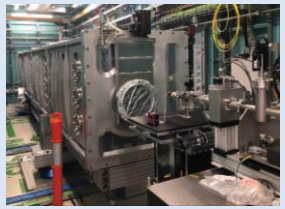
Micro-CT
(8 – 40 keV)
Sept 2022



MEX-1
(3.5 - 13.6 keV)
Nov 2022



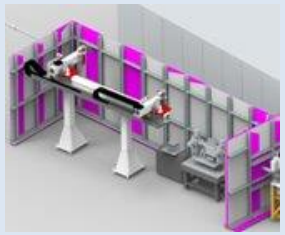
MEX-2
(1.3 - 3.5 keV)
Mid-2023



BioSAXS
(8 – 15 keV)
Mid-2023

MX3
(10-15 keV)
Mid-2024

ADS-1
(50 – 150 keV)
Late-2024



ADS-2
(45 – 90 keV)
Late-2024

NANO
(5 – 18 keV)
Mid-2025

The Scientific Computing Team

Scientific Computing founded in June 2017

Support Science and Users

- Experiment Control
- Data Acquisition
- Data Processing
- Data Analysis

Our Team

- 1 manager
- 19 members
  - 1 principal engineer
  - 10 PhDs
  - 47% gender split

Beamline Groups

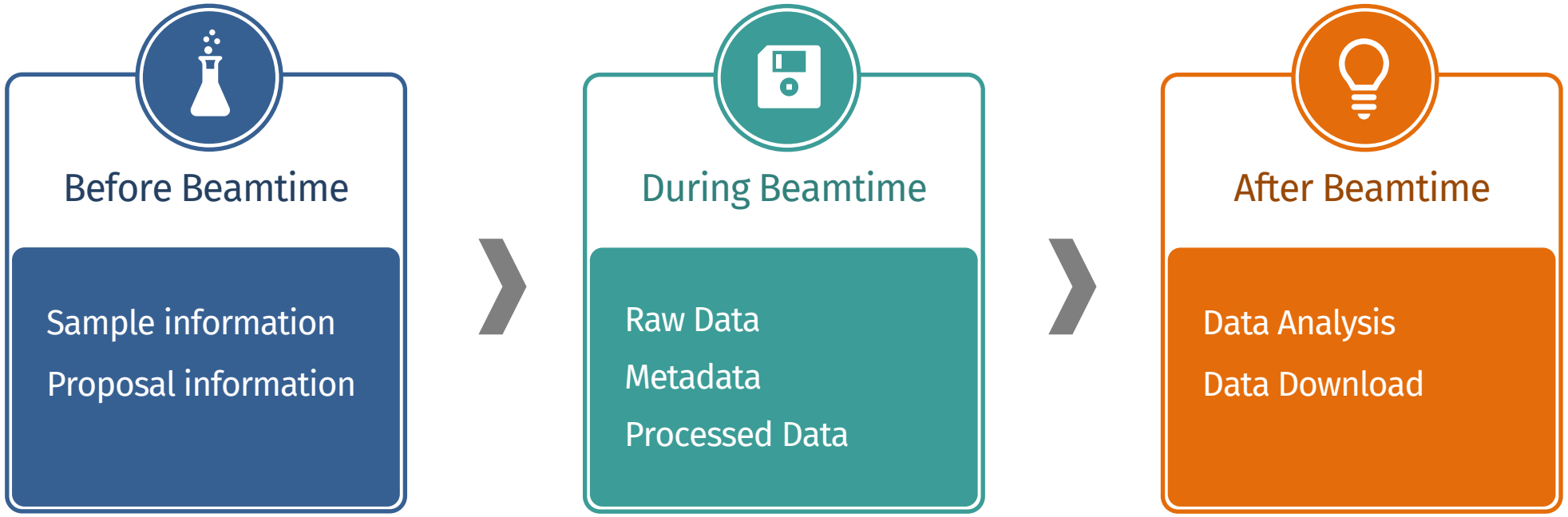
- Imaging
- Spectroscopy
- Scattering
- Crystallography
- Diffraction
- Microscopy

Cross-functional teams
(3 to 4 members)

- Hardware control
- Backend
- Frontend
- Processing

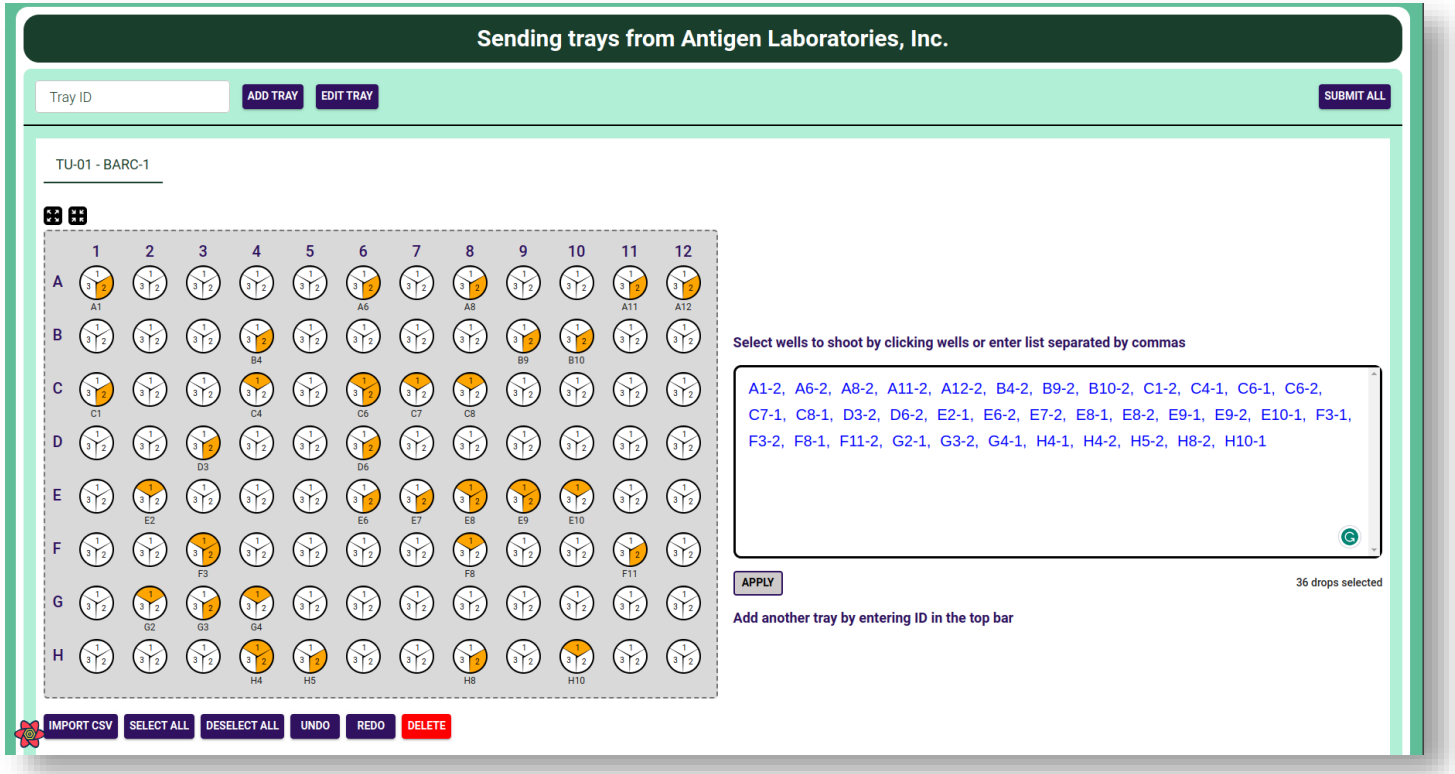
ANSTO

The Data Journey



Sample Management

Example: MX3 tray screening interface



Management

- Implementation depends on beamline
- Users enter sample details prior to experiment
- Provides crucial metadata

Typical Features

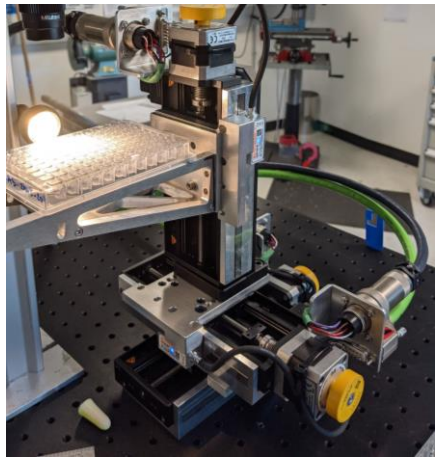
- Generate shipping labels
- Assign samples to sample holder
- Associate sample with experiment and data

python™ FastAPI PostgreSQL

ANSTO

Data Collection Architecture

Hardware



Run Engine

bluesky

Metadata

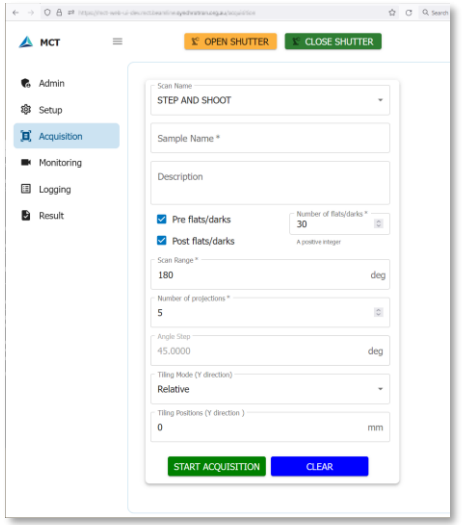
kafka

Data Streaming

ØMQ

redis

python™ bluesky



Data Collection

FastAPI

python™

Local Storage

HDF

Data Product

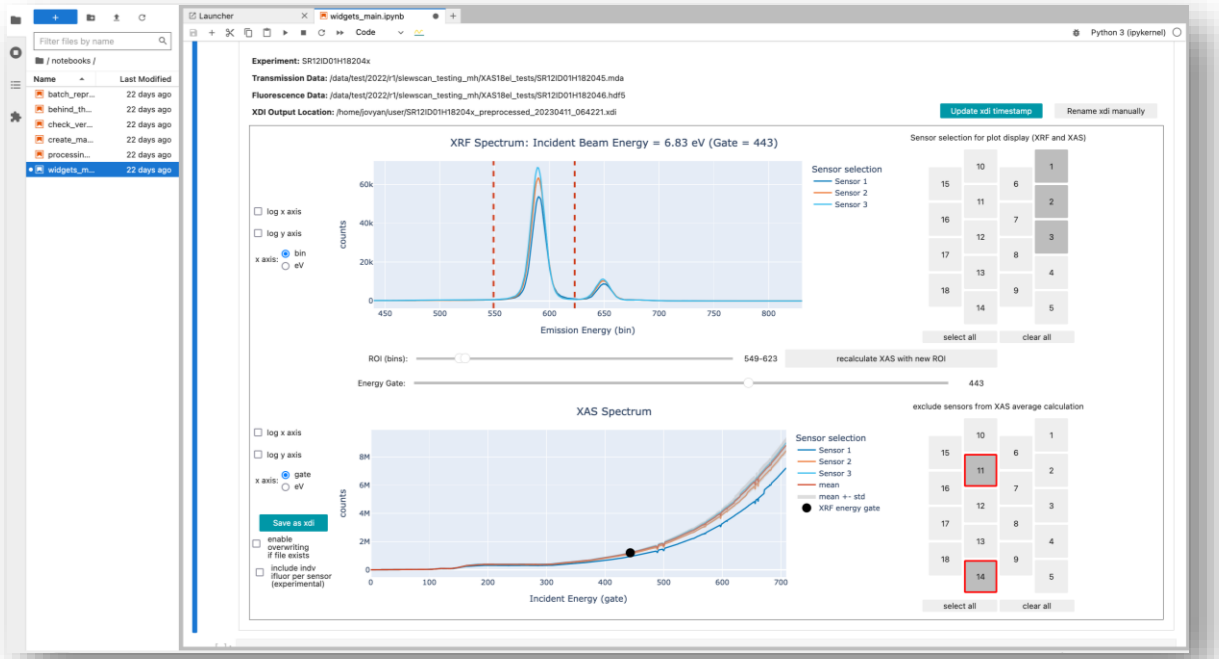
ANSTO

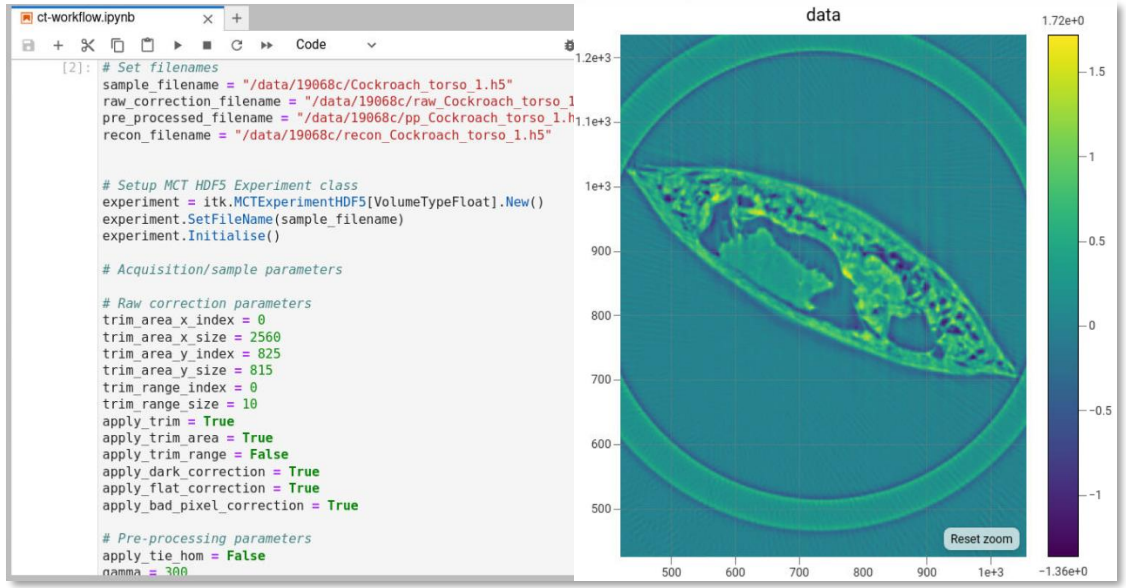
Data Processing

High performance computing

- Computing system
  - Comprised of 50+ physical servers
  - GPU nodes for image processing

- Beamline specific implementation
  - Heavily depends on beamline
  - Off-the-shelf and in-house tools
  - Heavy use of frameworks (e.g. ITK, Prefect)
  - Runs in Docker containers on Kubernetes

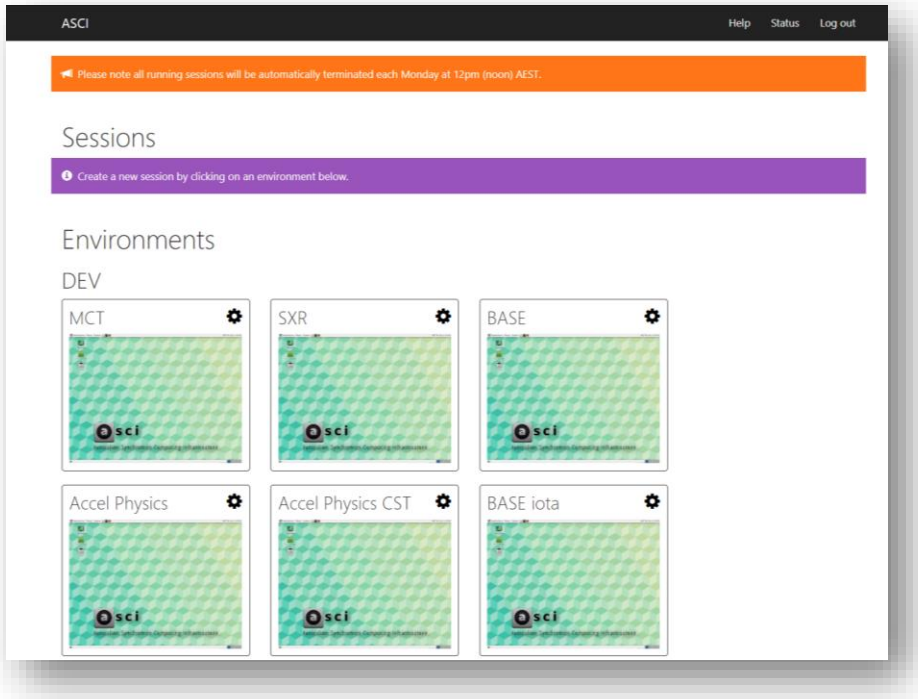




Data Analysis & Visualisation (ASCI)

Remote analysis platform

- Remote desktop environment in browser
- User starts session with tools pre-installed and data mounted

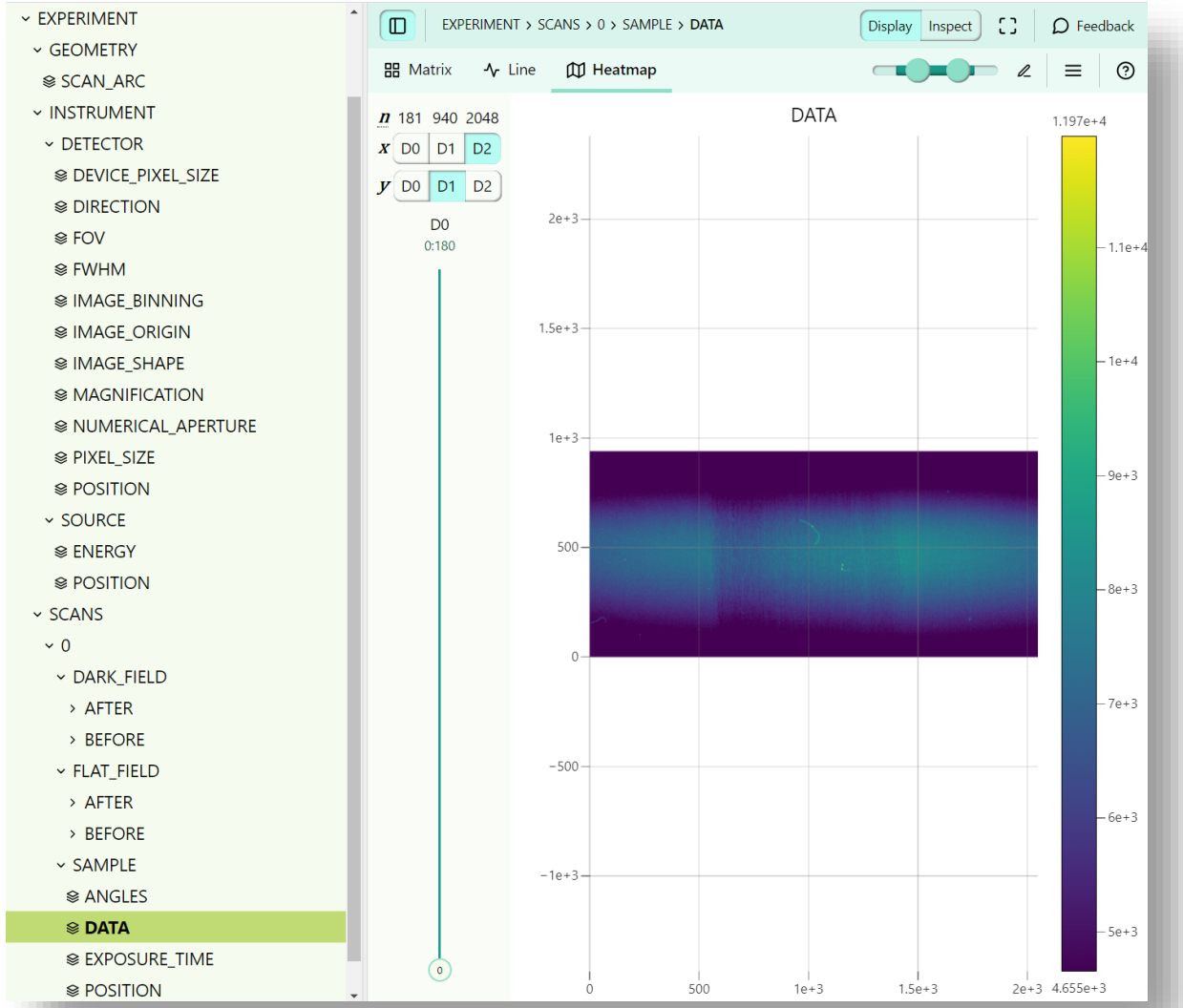


docker

python™

ANSTO

The Data Product



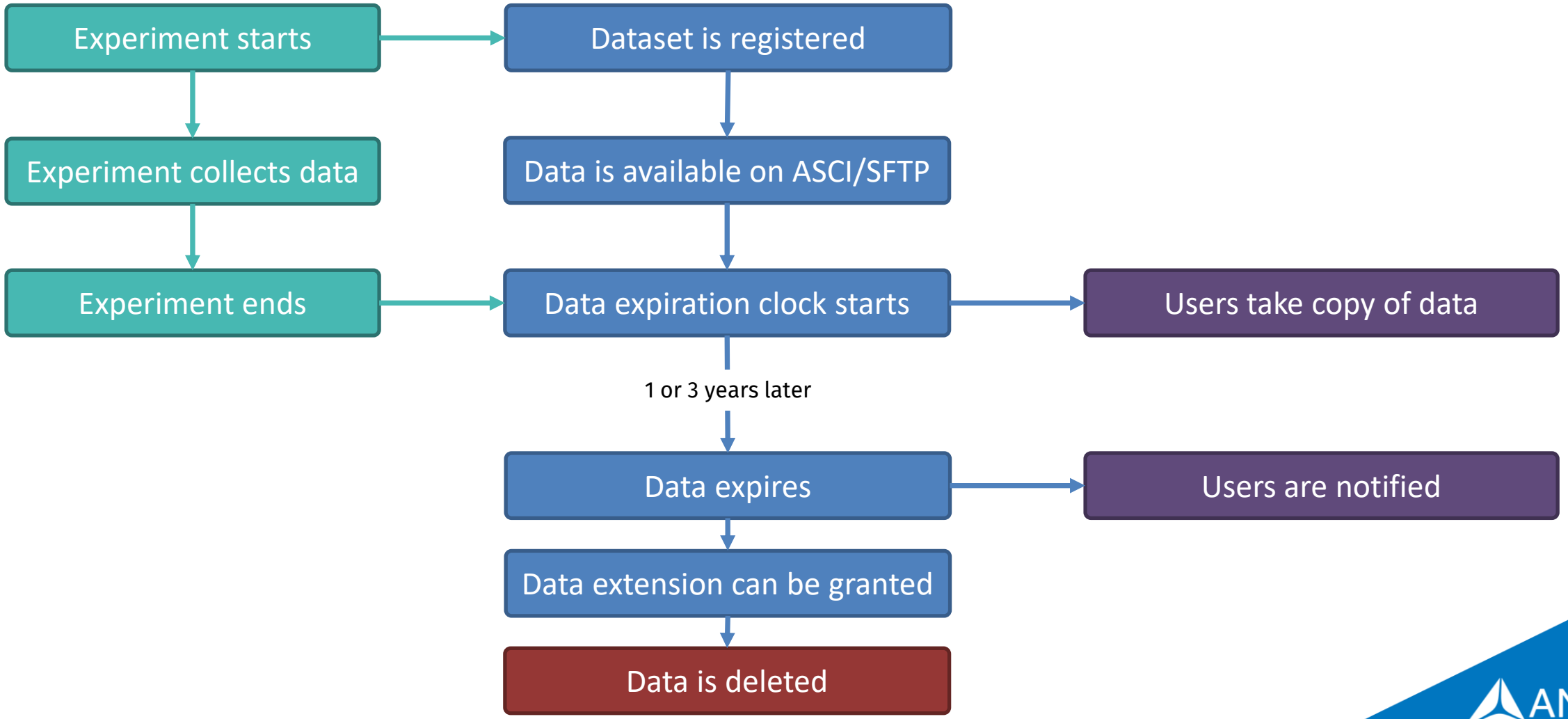
Concept of Data Product

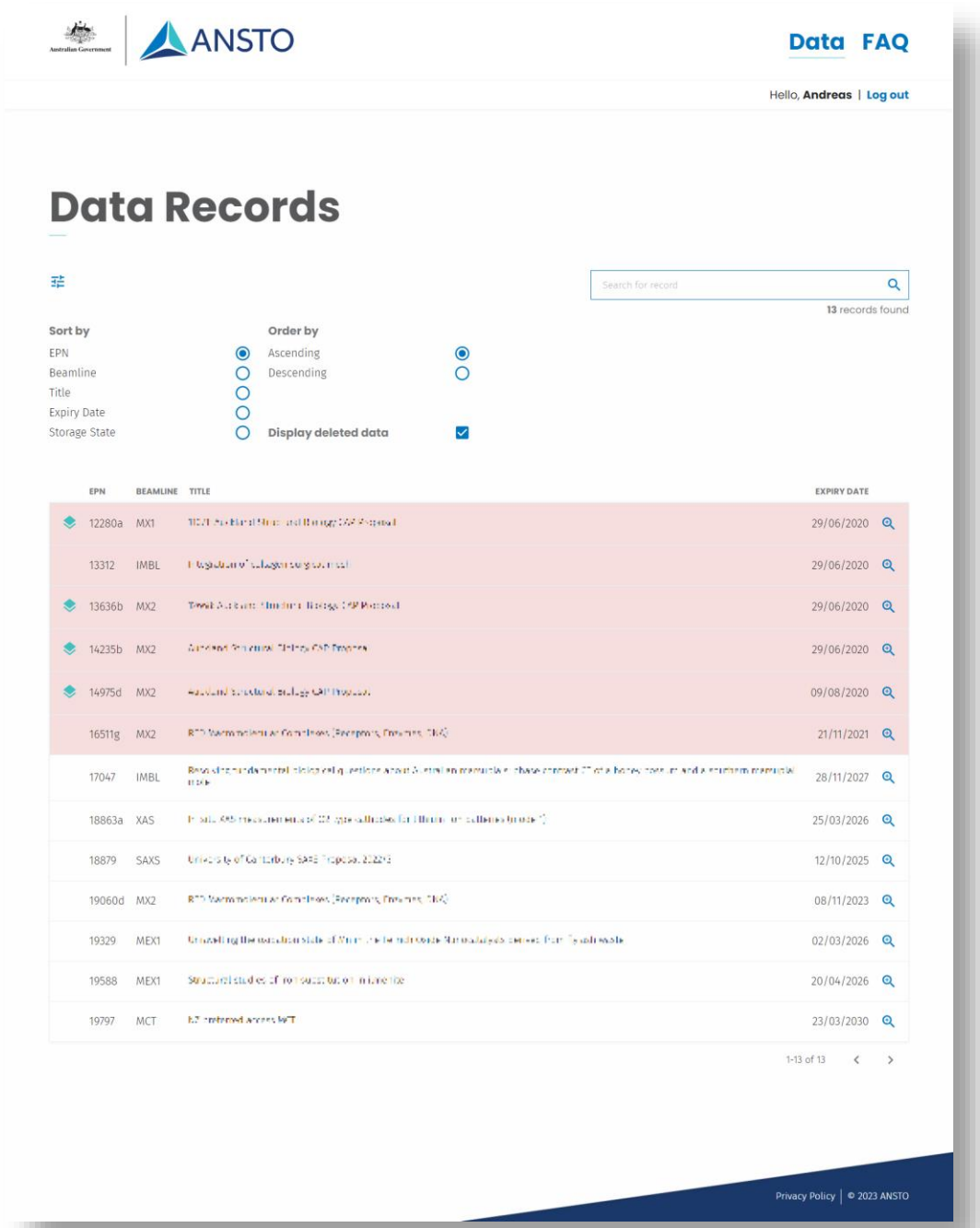
- Raw + Metadata in as few files as possible
- Target users

HDF5 is our standard container

- Schema depends on Beamline
  - MX beamline use NeXus NXmx
  - MCT uses custom format inspired by NeXus
  - Other beamlines still in discussion
- Defined through models
  - H5Pydantic Python library under development

Data Lifecycle





Data Page for Users

- Users can check the state of their datasets
  - Shows expiration dates, title, EPN, data size
  - Can sort by various criteria
- Includes FAQ
  - Explains website
  - Explains lifecycle

ANSTO

Summary

Data Retention

- Users responsible for long term data storage
- Local copy of data deleted after 1 or 3 years

Focus on Data Product

- Described and verifiable
- Rich Metadata
- Informed by processing and analysis tools
- Follow community standards where possible

ANSTO

Questions?

ANSTO

Data Retention Policy

Guidelines

This document outlines the conditions of access to and retention of raw and other experimental data by Users at ANSTO Clayton Campus; the site of the Australian Synchrotron.

Data Strategy

Users of the Australian Synchrotron shall have access to raw and reduced experimental data, and relevant metadata collected during an experiment (collectively 'data') under the following conditions:

(a) It is the User's responsibility to ensure that they take a copy of the data collected during an experiment.

(b) ANSTO will apply "best-efforts" for data access, retention and security; however, ANSTO makes no guarantees to retain or protect data collected from experiments.

(c) ANSTO will endeavour to retain data collected from experiments, *for up to 3 years for low data-rate beamlines*, or *for 12 months in the case of high data-rate beamlines.*

(d) Users may bring portable storage devices to the beamlines to download the data collected from experiments to these devices. ANSTO may also support the downloading of data via remote access.

(e) ANSTO will endeavour to provide remote data access capability for Users to download data collected during experiments to a machine of their choice. Currently, this capability is provided through a web interface however ANSTO may change how such services are provided at its discretion.

(f) ANSTO may restrict access to data to Users that are either named on the Experiment Authorisation form for the experiment, or as modified by the Principal Investigator using an administrative process approved by ANSTO.

(g) If third party data storage systems are used, ANSTO accepts no responsibility for data security or integrity.

(h) When data has been archived for more than 3 years after its collection for low data-rate beamlines, or for 12 months in the case of high data-rate beamline experiments, it may be deposited in a public access archive, or deleted, at the discretion of ANSTO.

(i) Users must write to the Australian Synchrotron User Office if they do not wish data collected from experiments to be deposited in a public archive.

Data Access and Retention Guide

(j) If ANSTO intends to delete data from its Australian Synchrotron data store, it will make best efforts to inform the relevant User of this action one month prior to deletion. It will be the responsibility of the User to ensure that they have downloaded a copy of their data prior to deletion.

(k) The retention periods for data storage, and this policy, may be reviewed annually by ANSTO.

Definitions

Low data-rate beamlines - Any beamline at the Australian Synchrotron that generates modest data rates – typically less than one terabyte per experiment. At the time of producing this document, low data-rate beamlines include: IRM, PD, SAXS/WAXS, SXR, THz/Far-IR, XAS, XFM and offline instruments.

High data-rate beamline - Any beamline at the Australian Synchrotron that generates high data rates – typically more than one terabyte of data per experiment. At the time of producing this document, high data-rate beamlines include: IMBL, MX1 and MX2; however this may include other beamlines in the future.

IMBL – Imaging and Medical beamline

IRM – Infrared Microspectroscopy beamline

MX1 – Macromolecular Crystallography beamline

MX2 – Microcrystallography beamline

PD – Powder Diffraction beamline

SAXS/WAXS – Small and Wide Angle Scattering beamline

SXR – Soft X-ray Spectroscopy beamline

THz/Far-IR – Terahertz and Far-Infrared beamline

XAS – X-ray Absorption Spectroscopy beamline

XFM – X-ray Fluorescence Microscopy beamline

User – A researcher that uses the beamlines or other facilities at the Australian Synchrotron.