Towards a generalised approach for defining, organising and storing metadata from all experiments at the ESRF

by Andy Götz
ESRF

IUCR Satellite Workshop on Metadata
29th ECM (Rovinj) 2015

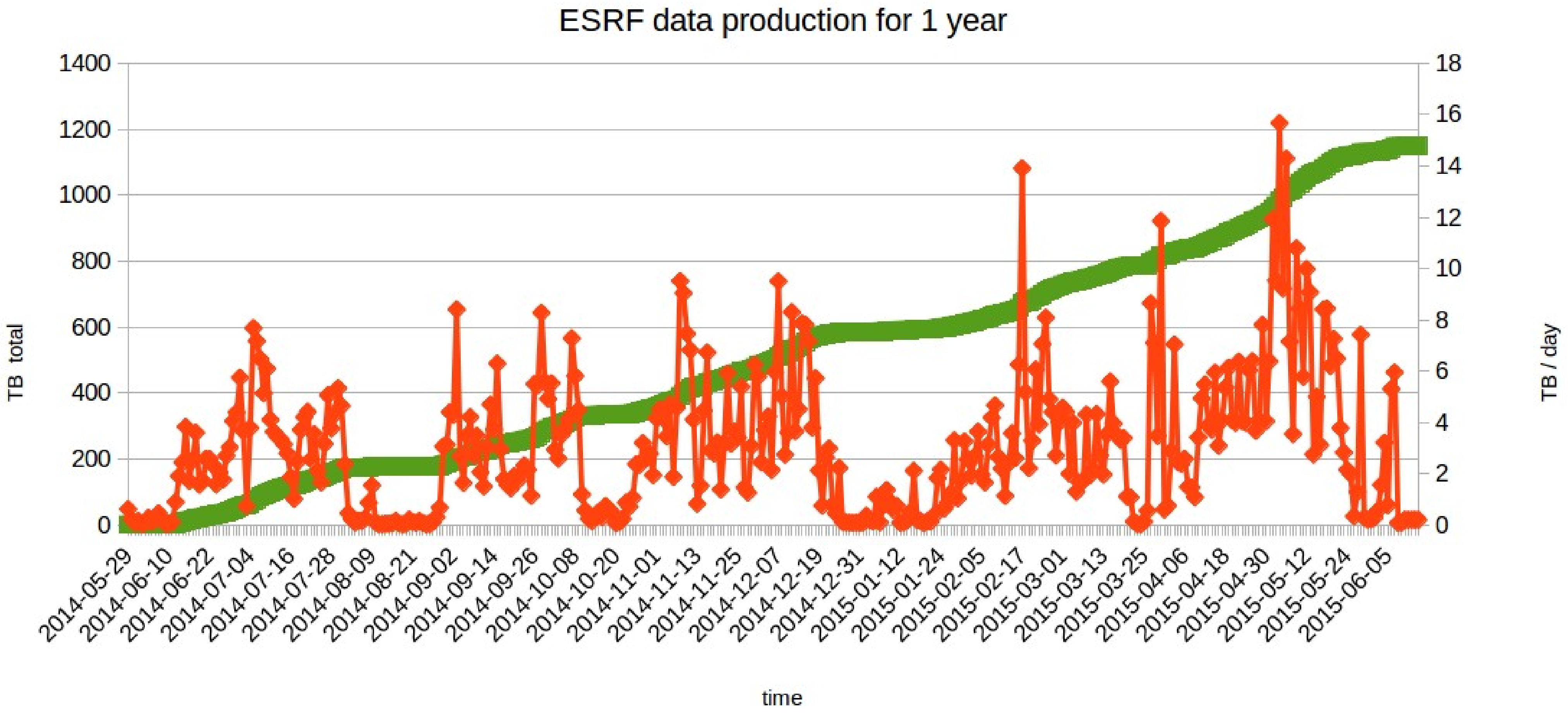
Looking towards the future

photo by **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

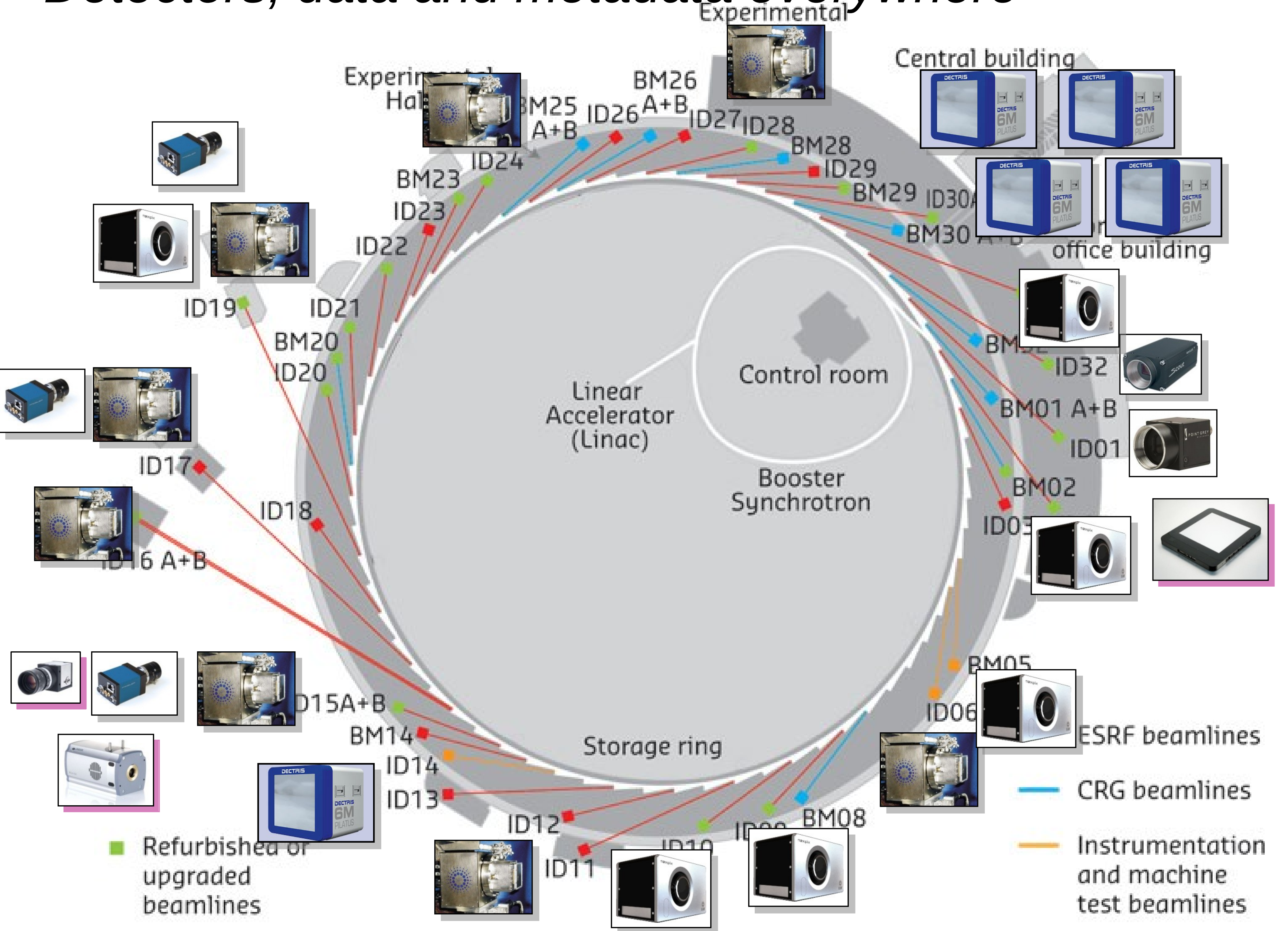
ESRF

- The European Synchrotron (Grenoble, France)
- 40+ Beamlines running 24/7
- Produced 1.2+ Petabytes of raw data in 2014-2015
- Metadata
  - Metadata well defined and managed for macromolecular crystallography (MX)
  - Non-unified approach for 35+ non-MX beamlines
- Upgrading source to a diffraction limited storage ring with 50+ more brilliance and coherence

ESRF Data Production



Detectors, data and metadata everywhere



New Detectors every year

copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

What do we mean by Metadata ?

- *“When talking about metadata we are often talking past each other”*

- *“Data about data doesn't mean anything”*

- In this talk we define **metadata** to mean:

  Data needed to reduce or analyse raw data i.e. in addition to the raw data

Metadata can also be data ...

alay). 5. Engaged in busine
scale. 6. Unpretentious; mod
ung. 8. Characterized by littlen
9. Belittled; humiliated. 10. Di
everages. 11. Soft; low: *a small v*
s: *Cut it up small.* 2. Softly. 3.
. Something smaller than the rest:
Plural. a. Small things collectively.
le English *smal(l)*, Old E
in Ap x. —**small′ness** *n.*
onyms: *sma little minutive, minute, minia*
cule, tiny, wee, p se adjectives describe
things whose physic is markedly below that
age. *Small* and *little* can often be used interchange
eral, *small* has the wider application; with referenc
size, *little* is usually more emphatic in implying sha
from the average. *Little* is sometimes used also to a
charm, endearment, or pathos to the term modifie
means very, often abnormally, small. *Minute* desc
small to the point of being difficult to see. *Miniat*

copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

Three classes of Metadata

copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

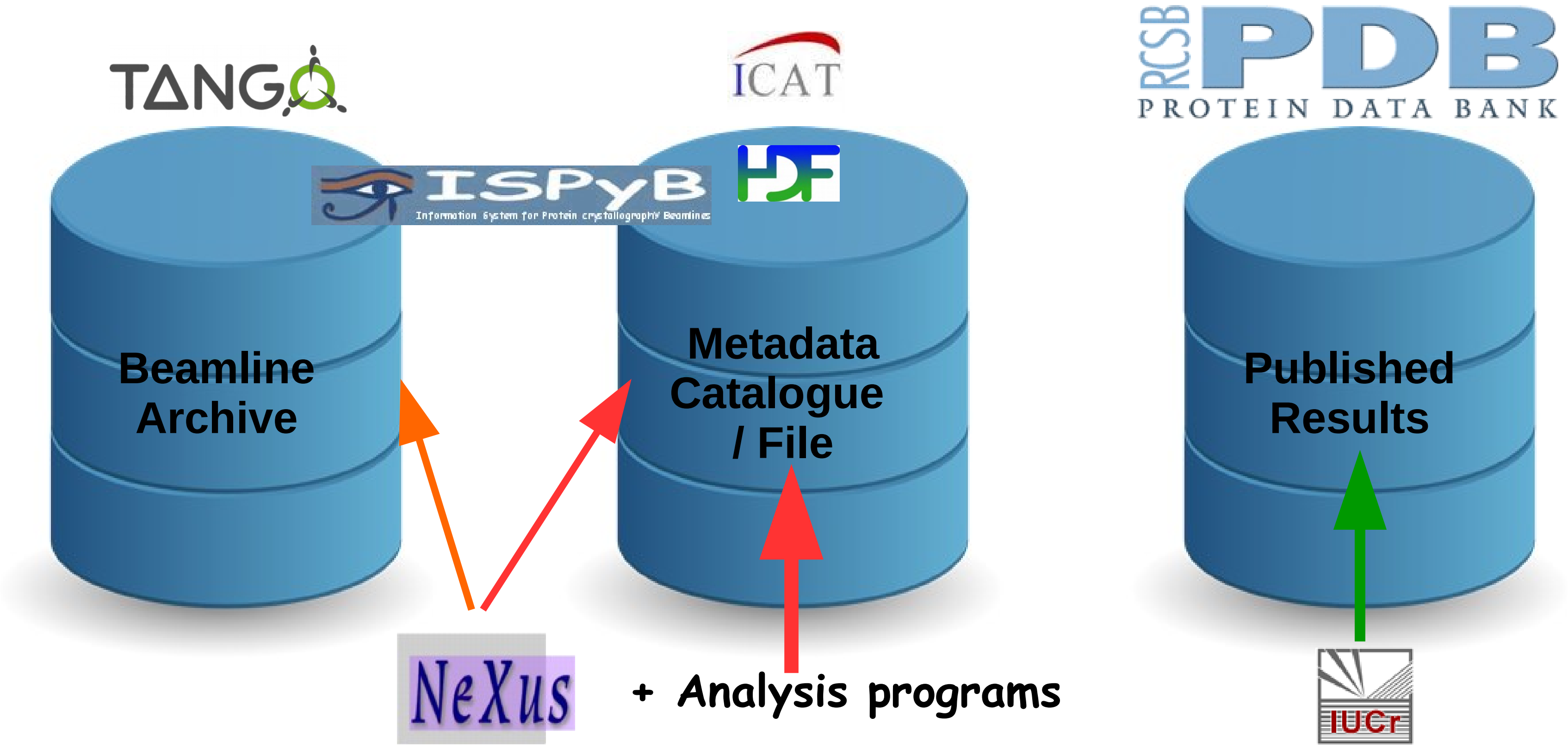
Three classes of Metadata

- **Beamline+Sample** – everything that describes the setup of the beamline + sample .
  - Used by beamline staff + experiment

➔ **Experiment** – everything that describes the experiment and how it was conducted
  - Used by data analysis programs

- **Results** – everything that describes the results of the analysis and the model
  - Used by scientists and journals

3 x Metadata Classes = 3 x Databases


TANGO
ICAT
RCSB PDB PROTEIN DATA BANK
ISPyB Information System for Protein crystallography Beamlines
HDF
NeXus
IUCr

Beamline Archive

Metadata Catalogue / File

Published Results

+ Analysis programs

Beamline + Sample

Experiment

Results

Why don't we manage Metadata well?

- **Single beamlines** lack the critical mass to drive forward metadata standardisation
- Beamlines use **multiple techniques**
- Same technique on multiple beamlines
- **Community standards** often don't exist
- **New techniques** being invented regularly
- **Exception to** the **rule** is when multiple beamlines using the same technique work together e.g. MX (see Gordon's talk)

New global approach @ ESRF

- Moving to a **global** site wide **solution** tailored to **local** needs
- **Commonly defined framework** for input and output
- Implement a **site wide** solution offering **same services** for all experiments
- **Constrain** the global and local definitions using **Nexus + HDF5 + icat**
- Give **priority** to metadata required for data **reduction** and **analysis programs**
- Address **all techniques** on **all beamlines**
- **Implement** a **metadata + data policy**

Goals for Metadata @ ESRF for ALL beamlines

1. Define metadata for all experimental techniques
2. Define data format(s) for automated data analysis
3. Annotate data with metadata and store in HDF5
4. Archive all metadata forever
5. Provide access to metadata
6. Implement DOI for data for provenance + publications
7. Provide users efficient download data service(s)
8. Archive (not-for-free) service to curate raw data
9. Implement the ESRF Data Policy (as soon as it has been defined by the management)

ICAT

Advanced Metadata management

- For **certain** techniques (MX, BIOSAXS, XRPD) :
  - LIMS* to track samples from the lab to publication
  - Track samples location for safety purpose
  - Advanced web interface for
    - preparing experiments,
    - running experiments and
    - displaying results
  - Upload results to public database(s)
  - Mail-in service for industrial users
  - Data Analysis As a Service (DAAS)

ISPyB
Information System for Protein crystallographY Beamlines

* Laboratory Information Management System

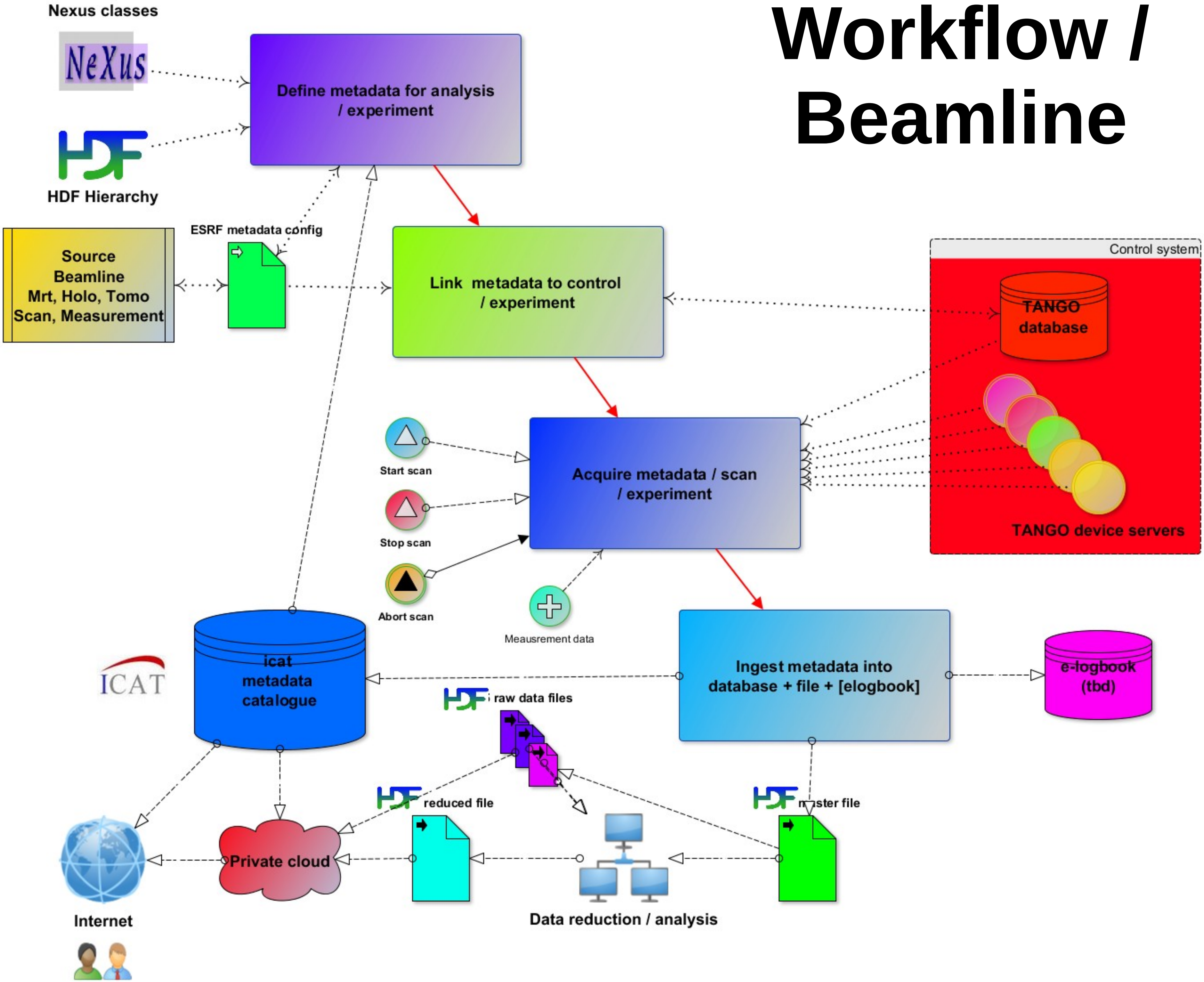
MX+ISPyB talk is tomorrow

copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

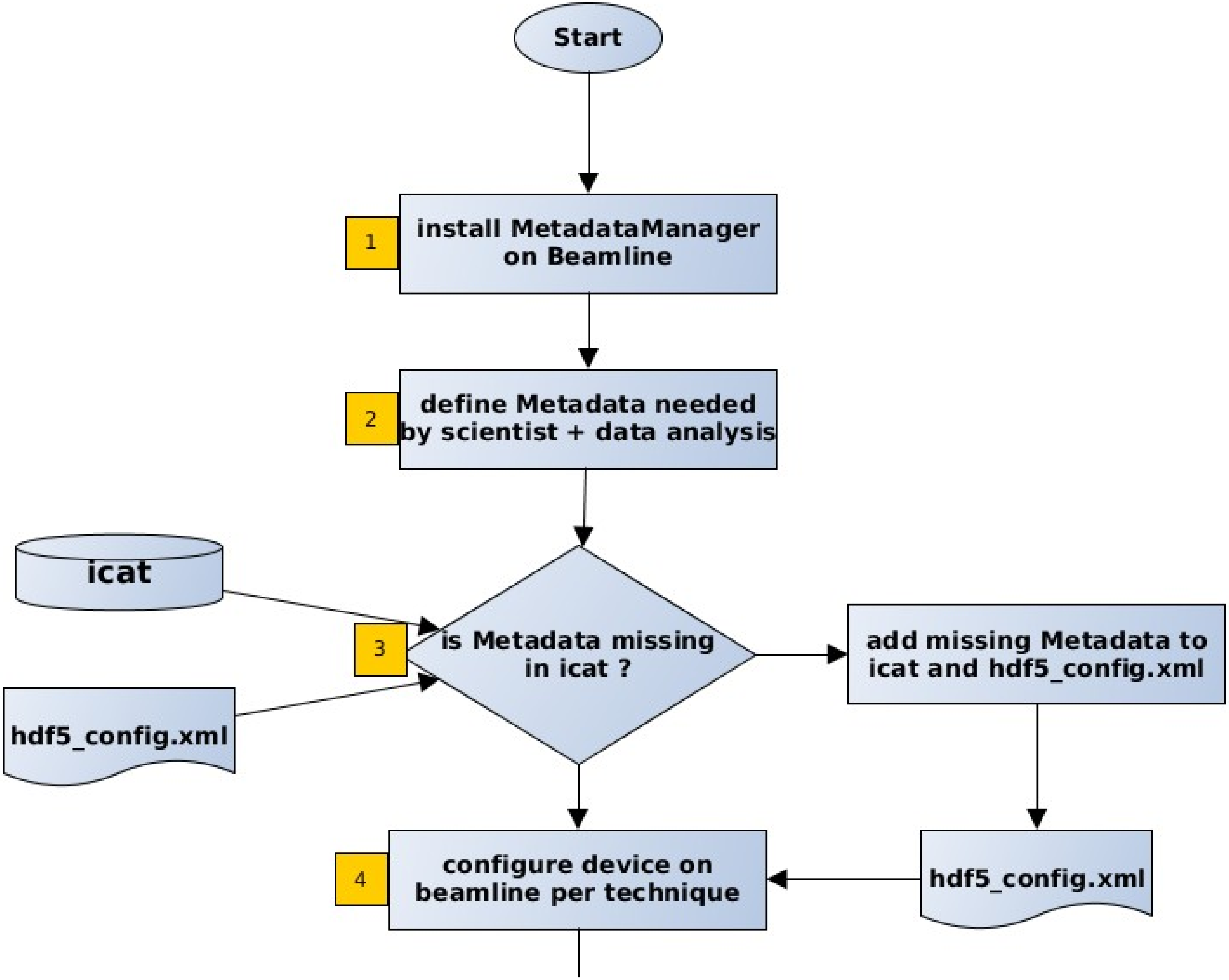
Global approach builds on existing standards

- **HDF5** – use the hierarchical structure to store data from multiple techniques in single file or master file + links to data files
- **Nexus** – use the Nexus classes as much as possible (do not reinvent the wheel)
- **Icat** – use icat for metadata catalogue and profit from all the services + the community
- **ISPyB** – use ISPyB for advanced metadata management
- **CIF** – use IUCr standards where they exist for publishing results

Workflow / Beamline



add a new beamline or experimental technique



Start

1 install MetadataManager on Beamline

2 define Metadata needed by scientist + data analysis

icat

hdf5_config.xml

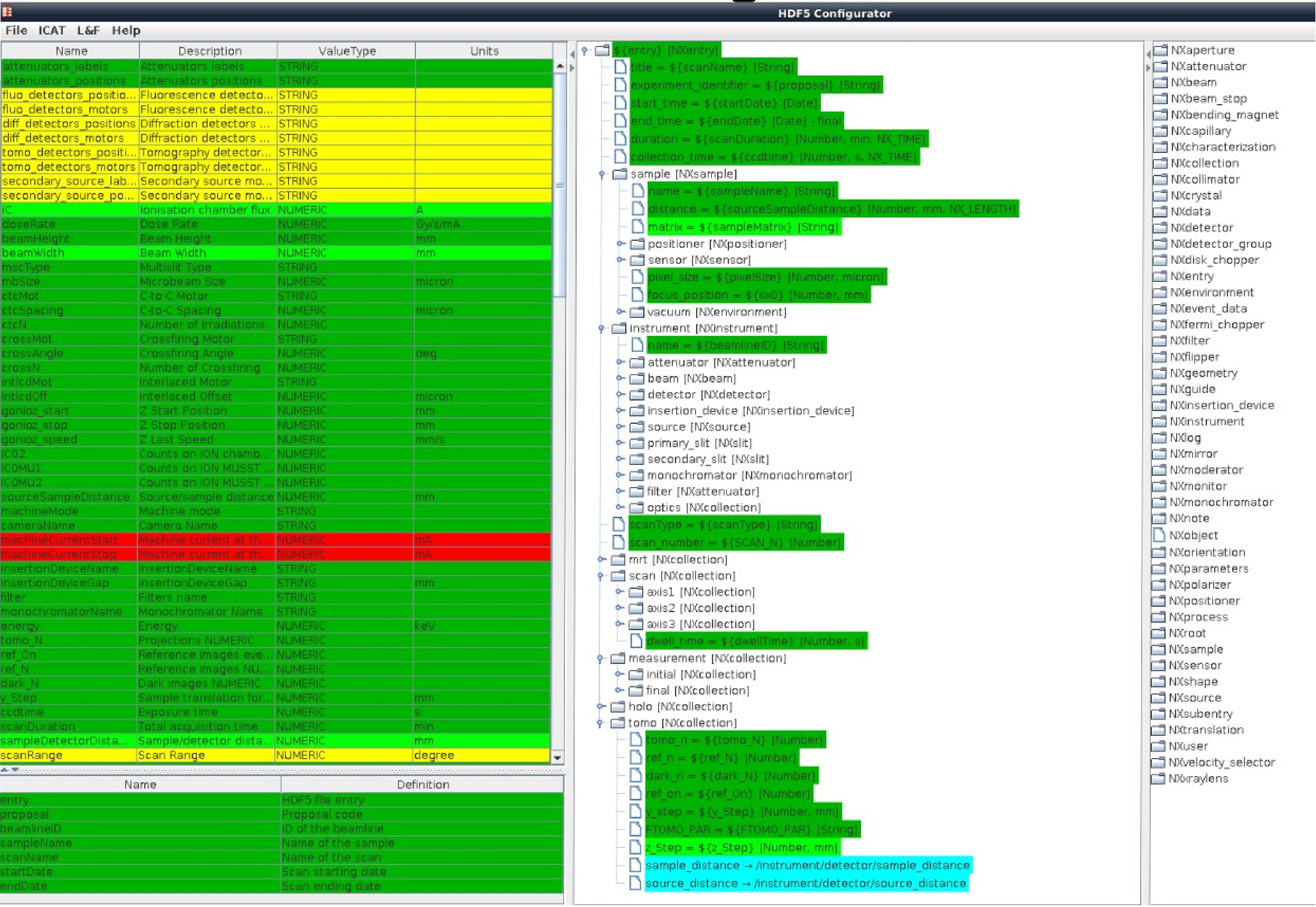
3 is Metadata missing in icat ?

add missing Metadata to icat and hdf5_config.xml

hdf5_config.xml

4 configure device on beamline per technique

HDF5 configurator



HDF5 Configurator

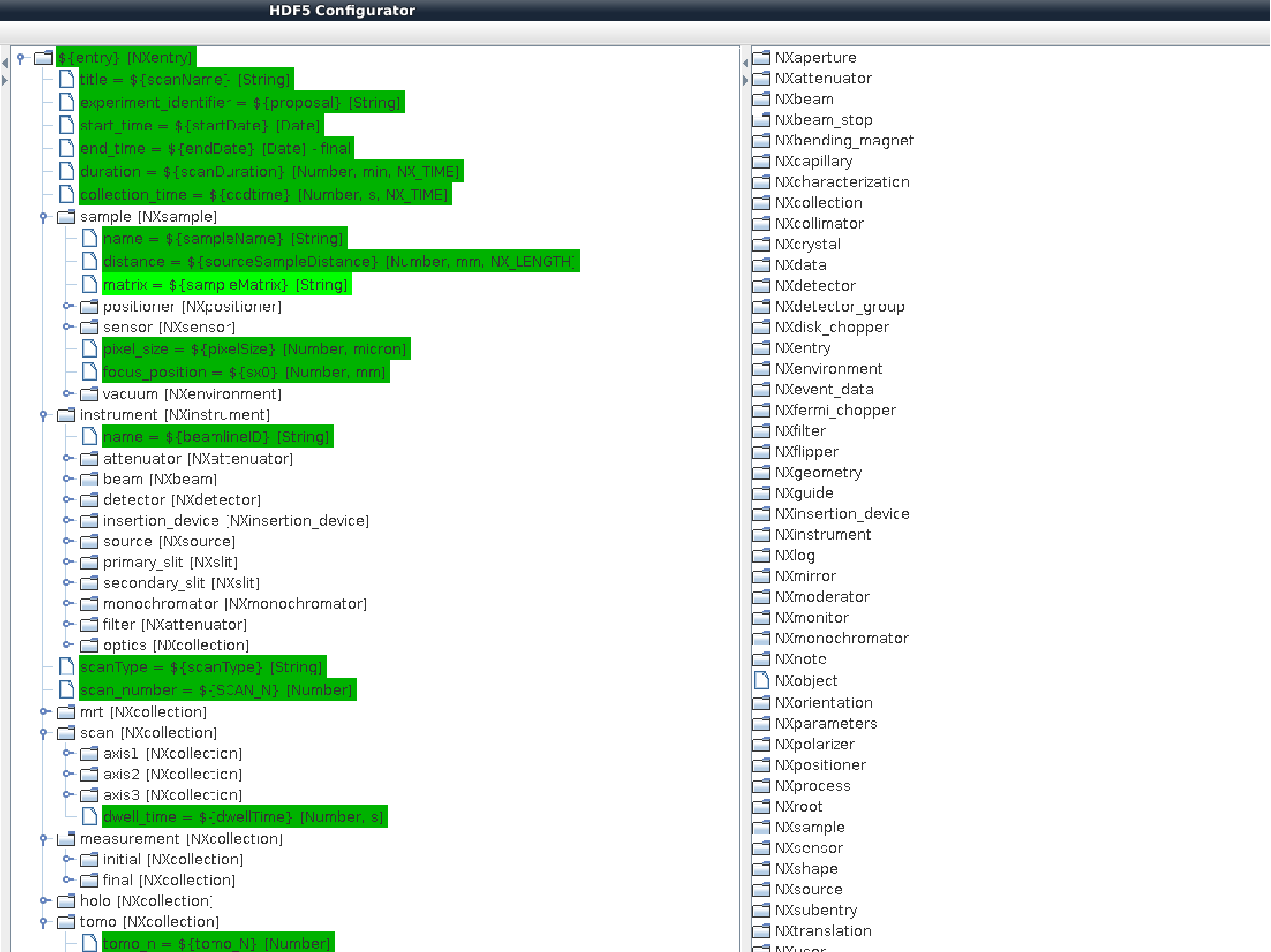
File ICAT L&F Help

Name	Description	ValueType	Units
attenuators_labels	Attenuators labels	STRING	
attenuators_positions	Attenuators positions	STRING	
fluo_detectors_positio...	Fluorescence detecto...	STRING	
fluo_detectors_motors	Fluorescence detecto...	STRING	
diff_detectors_positions	Diffraction detectors ...	STRING	
diff_detectors_motors	Diffraction detectors ...	STRING	
tomo_detectors_positi...	Tomography detector...	STRING	
tomo_detectors_motors	Tomography detector...	STRING	
secondary_source_lab...	Secondary source mo...	STRING	
secondary_source_po...	Secondary source mo...	STRING	
iC	Ionisation chamber flux	NUMERIC	A
doseRate	Dose Rate	NUMERIC	Gy/s/mA
beamHeight	Beam Height	NUMERIC	mm
beamWidth	Beam Width	NUMERIC	mm
mscType	Multislit Type	STRING	
mbSize	Microbeam Size	NUMERIC	micron
ctcMot	C-to-C Motor	STRING	
ctcSpacing	C-to-C Spacing	NUMERIC	micron
ctcN	Number of Irradiations	NUMERIC	
crossMot	Crossfiring Motor	STRING	
crossAngle	Crossfiring Angle	NUMERIC	deg
crossN	Number of Crossfiring	NUMERIC	
inticdMot	Interlaced Motor	STRING	
inticdOff	Interlaced Offset	NUMERIC	micron
gonioz_start	Z Start Position	NUMERIC	mm
gonioz_stop	Z Stop Position	NUMERIC	mm
gonioz_speed	Z Last Speed	NUMERIC	mm/s
IC02	Counts on ION chamb...	NUMERIC	
IC0MU1	Counts on ION MUSST ...	NUMERIC	
IC0MU2	Counts on ION MUSST ...	NUMERIC	
sourceSampleDistance	Source/sample distance	NUMERIC	mm
machineMode	Machine mode	STRING	
cameraName	Camera Name	STRING	
machineCurrentStart	Machine current at th...	NUMERIC	mA
machineCurrentStop	Machine current at th...	NUMERIC	mA
insertionDeviceName	insertionDeviceName	STRING	
insertionDeviceGap	insertionDeviceGap	STRING	mm
filter	Filters name	STRING	
monochromatorName	Monochromator Name	STRING	
energy	Energy	NUMERIC	keV
tomo_N	Projections NUMERIC	NUMERIC	
ref_On	Reference images eve...	NUMERIC	
ref_N	Reference images NU...	NUMERIC	
dark_N	Dark images NUMERIC	NUMERIC	
y_Step	Sample translation for...	NUMERIC	mm
ccdtime	Exposure time	NUMERIC	s
scanDuration	Total acquisition time	NUMERIC	min
sampleDetectorDista...	Sample/detector dista...	NUMERIC	mm
scanRange	Scan Range	NUMERIC	degree

Name	Definition
entry	HDF5 file entry
proposal	Proposal code
beamlineID	ID of the beamline
sampleName	Name of the sample
scanName	Name of the scan
startDate	Scan starting date
endDate	Scan ending date

```
${entry} [NXentry]
    title = ${scanName} [String]
    experiment_identifier = ${proposal} [String]
    start_time = ${startDate} [Date]
    end_time = ${endDate} [Date] - final
    duration = ${scanDuration} [Number, min, NX_TIME]
    collection_time = ${ccdtime} [Number, s, NX_TIME]
    sample [NXsample]
        name = ${sampleName} [String]
        distance = ${sourceSampleDistance} [Number, mm, NX_LENGTH]
        matrix = ${sampleMatrix} [String]
        positioner [NXpositioner]
        sensor [NXsensor]
        pixel_size = ${pixelSize} [Number, micron]
        focus_position = ${sx0} [Number, mm]
        vacuum [NXenvironment]
    instrument [NXinstrument]
        name = ${beamlineID} [String]
        attenuator [NXattenuator]
        beam [NXbeam]
        detector [NXdetector]
        insertion_device [NXinsertion_device]
        source [NXsource]
        primary_slit [NXslit]
        secondary_slit [NXslit]
        monochromator [NXmonochromator]
        filter [NXattenuator]
        optics [NXcollection]
    scanType = ${scanType} [String]
    scan_number = ${SCAN_N} [Number]
    mrt [NXcollection]
    scan [NXcollection]
        axis1 [NXcollection]
        axis2 [NXcollection]
        axis3 [NXcollection]
        dwell_time = ${dwellTime} [Number, s]
    measurement [NXcollection]
        initial [NXcollection]
        final [NXcollection]
    holo [NXcollection]
    tomo [NXcollection]
        tomo_n = ${tomo_N} [Number]
        ref_n = ${ref_N} [Number]
        dark_n = ${dark_N} [Number]
        ref_on = ${ref_On} [Number]
        y_step = ${y_Step} [Number, mm]
        FTOMO_PAR = ${FTOMO_PAR} [String]
        z_Step = ${z_Step} [Number, mm]
        sample_distance → /instrument/detector/sample_distance
        source_distance → /instrument/detector/source_distance
```

NXaperture
NXattenuator
NXbeam
NXbeam_stop
NXbending_magnet
NXcapillary
NXcharacterization
NXcollection
NXcollimator
NXcrystal
NXdata
NXdetector
NXdetector_group
NXdisk_chopper
NXentry
NXenvironment
NXevent_data
NXfermi_chopper
NXfilter
NXflipper
NXgeometry
NXguide
NXinsertion_device
NXinstrument
NXlog
NXmirror
NXmoderator
NXmonitor
NXmonochromator
NXnote
NXobject
NXorientation
NXparameters
NXpolarizer
NXpositioner
NXprocess
NXroot
NXsample
NXsensor
NXshape
NXsource
NXsubentry
NXtranslation
NXuser
NXvelocity_selector
NXxraylens

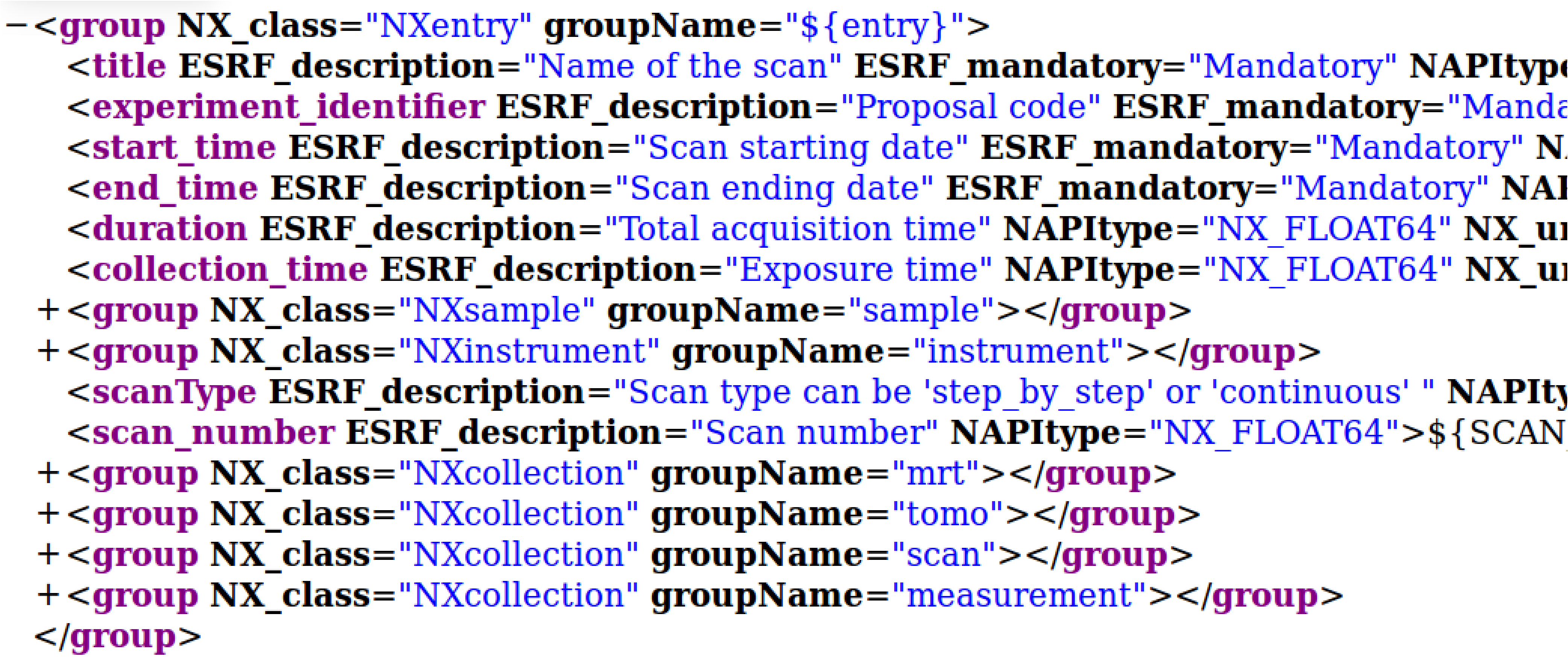


HDF5 Configurator

```
${entry} [NXentry]
    title = ${scanName} [String]
    experiment_identifier = ${proposal} [String]
    start_time = ${startDate} [Date]
    end_time = ${endDate} [Date] - final
    duration = ${scanDuration} [Number, min, NX_TIME]
    collection_time = ${ccdtime} [Number, s, NX_TIME]
    sample [NXsample]
        name = ${sampleName} [String]
        distance = ${sourceSampleDistance} [Number, mm, NX_LENGTH]
        matrix = ${sampleMatrix} [String]
        positioner [NXpositioner]
        sensor [NXsensor]
        pixel_size = ${pixelSize} [Number, micron]
        focus_position = ${sx0} [Number, mm]
        vacuum [NXenvironment]
    instrument [NXinstrument]
        name = ${beamlineID} [String]
        attenuator [NXattenuator]
        beam [NXbeam]
        detector [NXdetector]
        insertion_device [NXinsertion_device]
        source [NXsource]
        primary_slit [NXslit]
        secondary_slit [NXslit]
        monochromator [NXmonochromator]
        filter [NXattenuator]
        optics [NXcollection]
    scanType = ${scanType} [String]
    scan_number = ${SCAN_N} [Number]
    mrt [NXcollection]
    scan [NXcollection]
        axis1 [NXcollection]
        axis2 [NXcollection]
        axis3 [NXcollection]
        dwell_time = ${dwellTime} [Number, s]
    measurement [NXcollection]
        initial [NXcollection]
        final [NXcollection]
    holo [NXcollection]
    tomo [NXcollection]
        tomo_n = ${tomo_N} [Number]
```

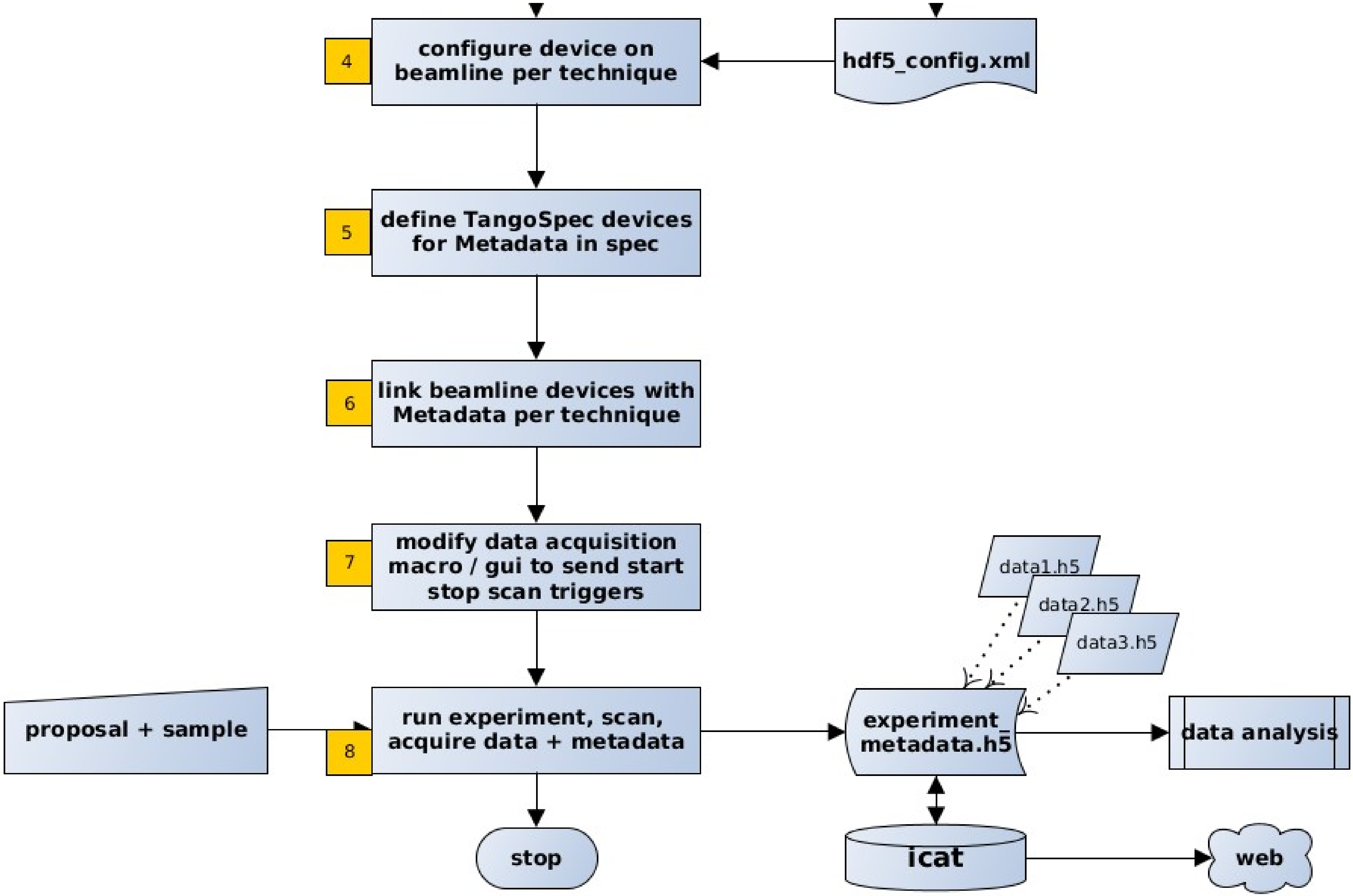
```
NXaperture
NXattenuator
NXbeam
NXbeam_stop
NXbending_magnet
NXcapillary
NXcharacterization
NXcollection
NXcollimator
NXcrystal
NXdata
NXdetector
NXdetector_group
NXdisk_chopper
NXentry
NXenvironment
NXevent_data
NXfermi_chopper
NXfilter
NXflipper
NXgeometry
NXguide
NXinsertion_device
NXinstrument
NXlog
NXmirror
NXmoderator
NXmonitor
NXmonochromator
NXnote
NXobject
NXorientation
NXparameters
NXpolarizer
NXpositioner
NXprocess
NXroot
NXsample
NXsensor
NXshape
NXsource
NXsubentry
NXtranslation
NXuser
```

Current global configuration of Nexus/HDF5 main classes



```
-<group NX_class="NXentry" groupName="${entry}">
   <title ESRF_description="Name of the scan" ESRF_mandatory="Mandatory" NAPItyp
   <experiment_identifier ESRF_description="Proposal code" ESRF_mandatory="Manda
   <start_time ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" N
   <end_time ESRF_description="Scan ending date" ESRF_mandatory="Mandatory" NA
   <duration ESRF_description="Total acquisition time" NAPItype="NX_FLOAT64" NX_u
   <collection_time ESRF_description="Exposure time" NAPItype="NX_FLOAT64" NX_u
 +<group NX_class="NXsample" groupName="sample"></group>
 +<group NX_class="NXinstrument" groupName="instrument"></group>
   <scanType ESRF_description="Scan type can be 'step_by_step' or 'continuous' " NAPIty
   <scan_number ESRF_description="Scan number" NAPItype="NX_FLOAT64">${SCAN
 +<group NX_class="NXcollection" groupName="mrt"></group>
 +<group NX_class="NXcollection" groupName="tomo"></group>
 +<group NX_class="NXcollection" groupName="scan"></group>
 +<group NX_class="NXcollection" groupName="measurement"></group>
 </group>
```

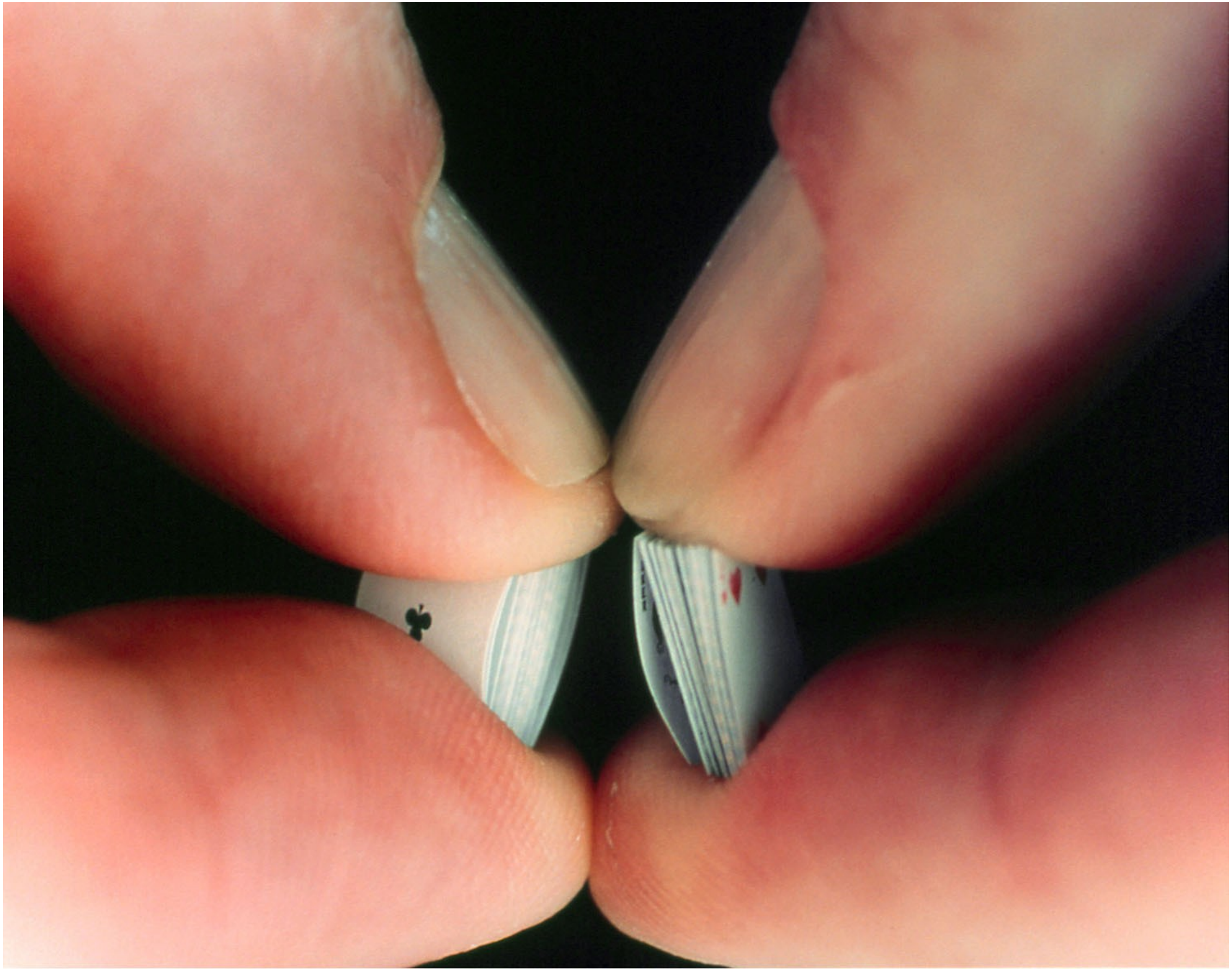
Definitions are a superset of all metadata required by all beamlines – a single beamline / experiment uses a subset of the information – multiple techniques stored in same file


4
configure device on beamline per technique
hdf5_config.xml
5
define TangoSpec devices for Metadata in spec
6
link beamline devices with Metadata per technique
7
modify data acquisition macro / gui to send start stop scan triggers
data1.h5
data2.h5
data3.h5
proposal + sample
8
run experiment, scan, acquire data + metadata
experiment_ metadata.h5
data analysis
stop
icat
web

Scan is a key concept

- The notion of a scan is **fundamental** to the experiment and data analysis programs
- A scan represents a **dataset**. It is comprised of one or more data acquisitions runs with zero or more changing parameters (scanning)
- A scan is a **meta-concept** for grouping data
- A scan can be made up of **sub-scans**

Flexible metadata definitions



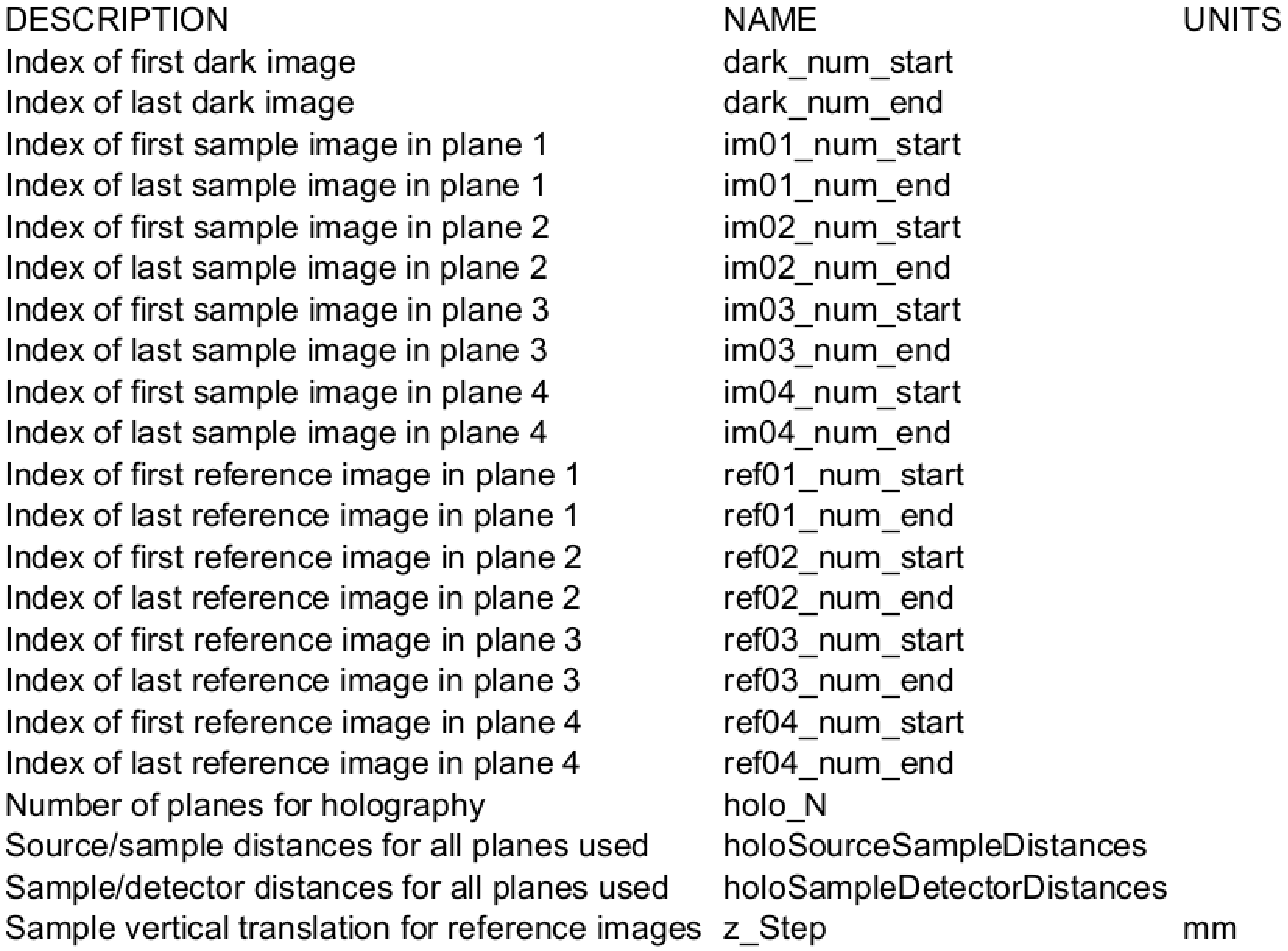
copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

Flexible Metadata definitions

- Storing **metadata** with **images** is **not enough**
- Beamlines need to **change definitions** every time a **new technique** is added
- **Techniques** can be very **varied** and can be **beamline specific**
- More and more beamlines use **multiple techniques**
- Cannot expect international standards on all techniques

Typical request from a beamline

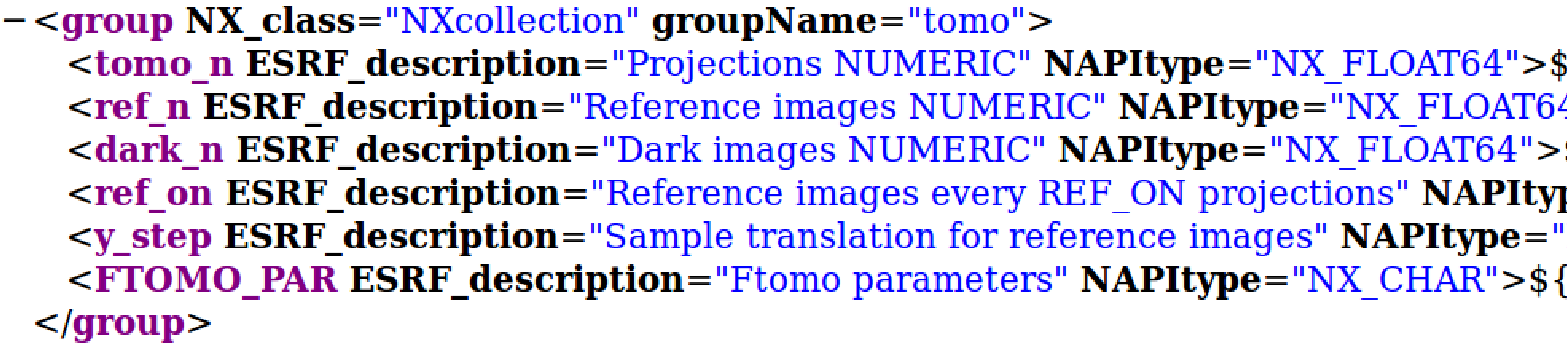
- **Holo-tomography**



DESCRIPTION	NAME	UNITS
Index of first dark image	dark_num_start	
Index of last dark image	dark_num_end	
Index of first sample image in plane 1	im01_num_start	
Index of last sample image in plane 1	im01_num_end	
Index of first sample image in plane 2	im02_num_start	
Index of last sample image in plane 2	im02_num_end	
Index of first sample image in plane 3	im03_num_start	
Index of last sample image in plane 3	im03_num_end	
Index of first sample image in plane 4	im04_num_start	
Index of last sample image in plane 4	im04_num_end	
Index of first reference image in plane 1	ref01_num_start	
Index of last reference image in plane 1	ref01_num_end	
Index of first reference image in plane 2	ref02_num_start	
Index of last reference image in plane 2	ref02_num_end	
Index of first reference image in plane 3	ref03_num_start	
Index of last reference image in plane 3	ref03_num_end	
Index of first reference image in plane 4	ref04_num_start	
Index of last reference image in plane 4	ref04_num_end	
Number of planes for holography	holo_N	
Source/sample distances for all planes used	holoSourceSampleDistances	
Sample/detector distances for all planes used	holoSampleDetectorDistances	
Sample vertical translation for reference images	z_Step	mm

Tomography

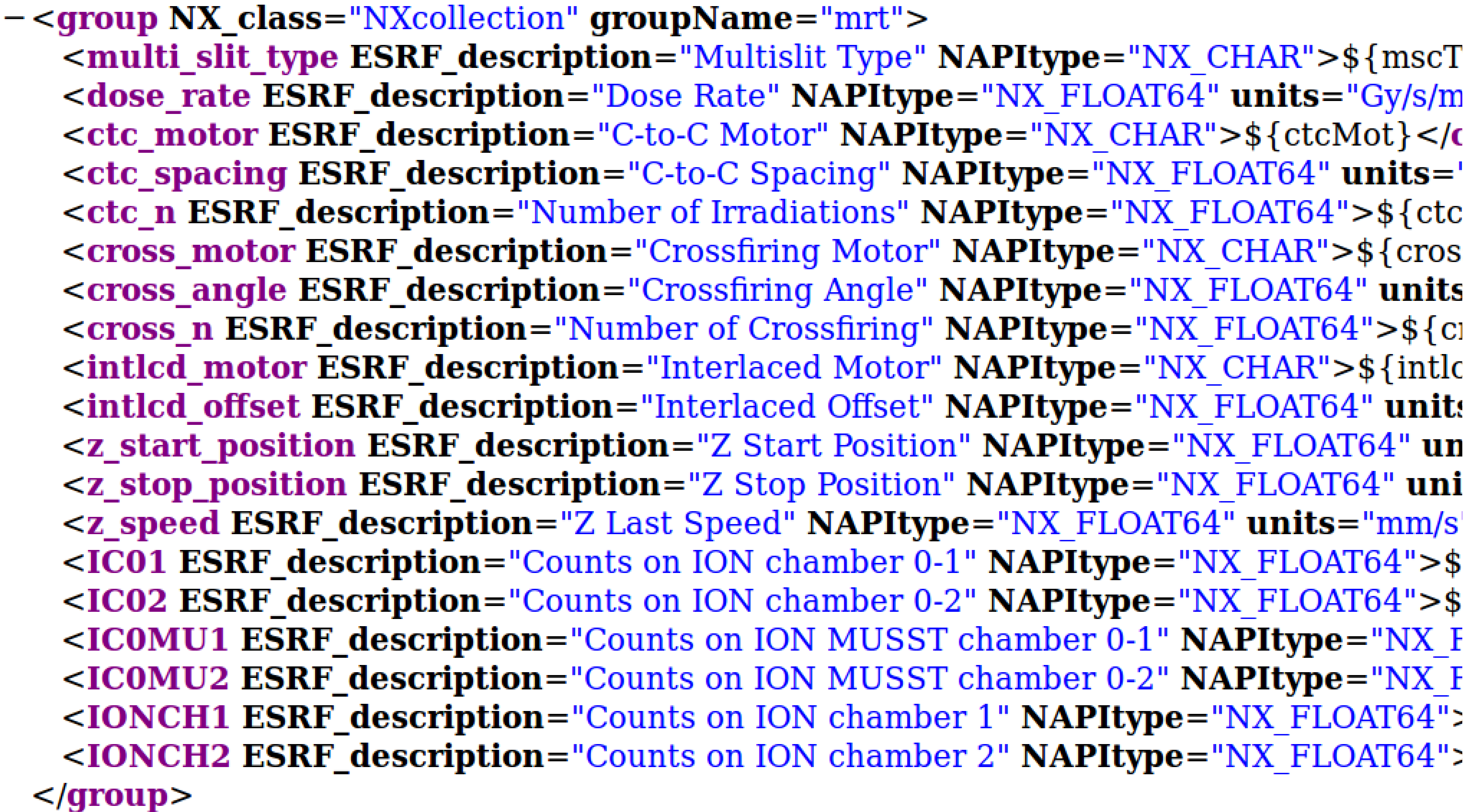
- Metadata required by tomography analysis programs



```
- <group NX_class="NXcollection" groupName="tomo">
    <tomo_n ESRF_description="Projections NUMERIC" NAPItype="NX_FLOAT64">$
    <ref_n ESRF_description="Reference images NUMERIC" NAPItype="NX_FLOAT6
    <dark_n ESRF_description="Dark images NUMERIC" NAPItype="NX_FLOAT64">
    <ref_on ESRF_description="Reference images every REF_ON projections" NAPIty
    <y_step ESRF_description="Sample translation for reference images" NAPItype="
    <FTOMO_PAR ESRF_description="Ftomo parameters" NAPItype="NX_CHAR">${
  </group>
```

Microbeam Radiation Therapy

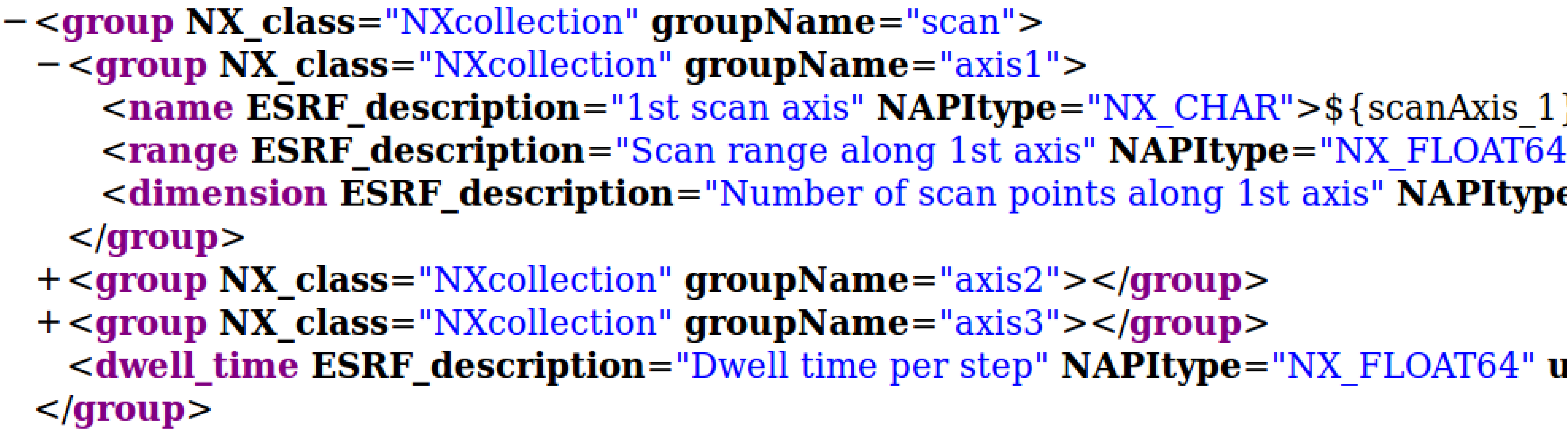
- Metadata required by MRT protocol



```
−<group NX_class="NXcollection" groupName="mrt">
   <multi_slit_type ESRF_description="Multislit Type" NAPItype="NX_CHAR">${mscT
   <dose_rate ESRF_description="Dose Rate" NAPItype="NX_FLOAT64" units="Gy/s/m
   <ctc_motor ESRF_description="C-to-C Motor" NAPItype="NX_CHAR">${ctcMot}</c
   <ctc_spacing ESRF_description="C-to-C Spacing" NAPItype="NX_FLOAT64" units="
   <ctc_n ESRF_description="Number of Irradiations" NAPItype="NX_FLOAT64">${ctc
   <cross_motor ESRF_description="Crossfiring Motor" NAPItype="NX_CHAR">${cros
   <cross_angle ESRF_description="Crossfiring Angle" NAPItype="NX_FLOAT64" units
   <cross_n ESRF_description="Number of Crossfiring" NAPItype="NX_FLOAT64">${cr
   <intlcd_motor ESRF_description="Interlaced Motor" NAPItype="NX_CHAR">${intlc
   <intlcd_offset ESRF_description="Interlaced Offset" NAPItype="NX_FLOAT64" units
   <z_start_position ESRF_description="Z Start Position" NAPItype="NX_FLOAT64" un
   <z_stop_position ESRF_description="Z Stop Position" NAPItype="NX_FLOAT64" uni
   <z_speed ESRF_description="Z Last Speed" NAPItype="NX_FLOAT64" units="mm/s"
   <IC01 ESRF_description="Counts on ION chamber 0-1" NAPItype="NX_FLOAT64">$
   <IC02 ESRF_description="Counts on ION chamber 0-2" NAPItype="NX_FLOAT64">$
   <IC0MU1 ESRF_description="Counts on ION MUSST chamber 0-1" NAPItype="NX_F
   <IC0MU2 ESRF_description="Counts on ION MUSST chamber 0-2" NAPItype="NX_F
   <IONCH1 ESRF_description="Counts on ION chamber 1" NAPItype="NX_FLOAT64">
   <IONCH2 ESRF_description="Counts on ION chamber 2" NAPItype="NX_FLOAT64">
 </group>
```

Generic Scan

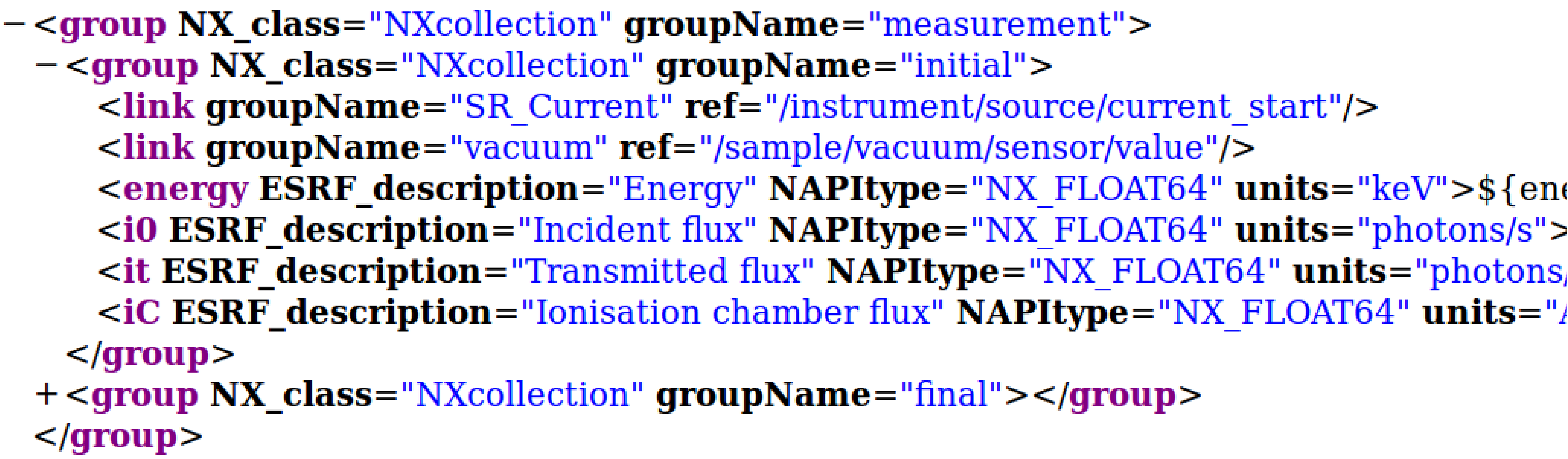
- Can be used for any technique e.g. fluorescence



```
−<group NX_class="NXcollection" groupName="scan">
  −<group NX_class="NXcollection" groupName="axis1">
      <name ESRF_description="1st scan axis" NAPItype="NX_CHAR">${scanAxis_1
      <range ESRF_description="Scan range along 1st axis" NAPItype="NX_FLOAT64
      <dimension ESRF_description="Number of scan points along 1st axis" NAPItype
    </group>
  +<group NX_class="NXcollection" groupName="axis2"></group>
  +<group NX_class="NXcollection" groupName="axis3"></group>
    <dwell_time ESRF_description="Dwell time per step" NAPItype="NX_FLOAT64" u
  </group>
```

Measurement

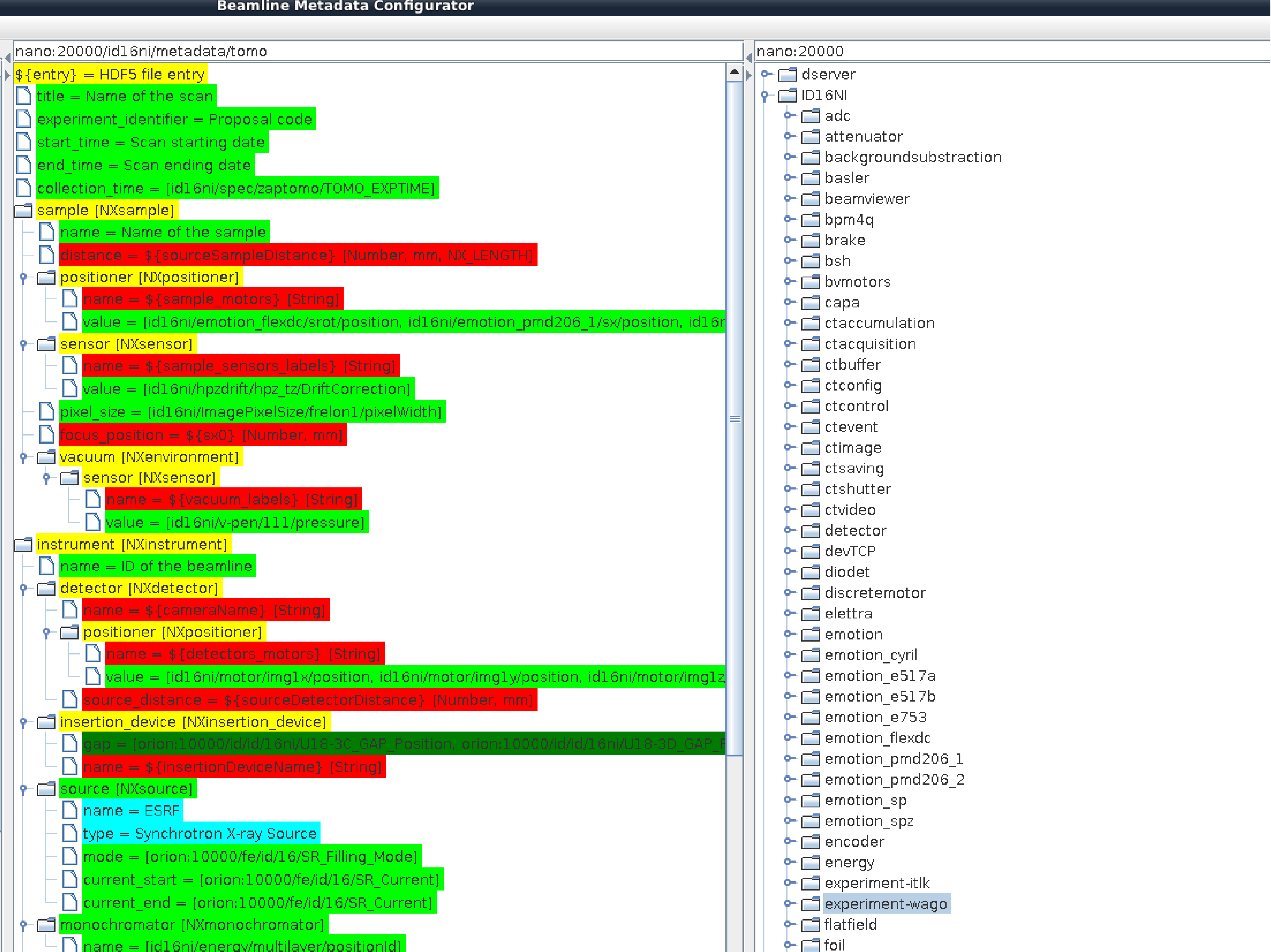
- Dynamically measured metadata



```
−<group NX_class="NXcollection" groupName="measurement">
  −<group NX_class="NXcollection" groupName="initial">
      <link groupName="SR_Current" ref="/instrument/source/current_start"/>
      <link groupName="vacuum" ref="/sample/vacuum/sensor/value"/>
      <energy ESRF_description="Energy" NAPItype="NX_FLOAT64" units="keV">${ene
      <i0 ESRF_description="Incident flux" NAPItype="NX_FLOAT64" units="photons/s">
      <it ESRF_description="Transmitted flux" NAPItype="NX_FLOAT64" units="photons/
      <iC ESRF_description="Ionisation chamber flux" NAPItype="NX_FLOAT64" units="
    </group>
  +<group NX_class="NXcollection" groupName="final"></group>
  </group>
```

Collecting Metadata on beamline

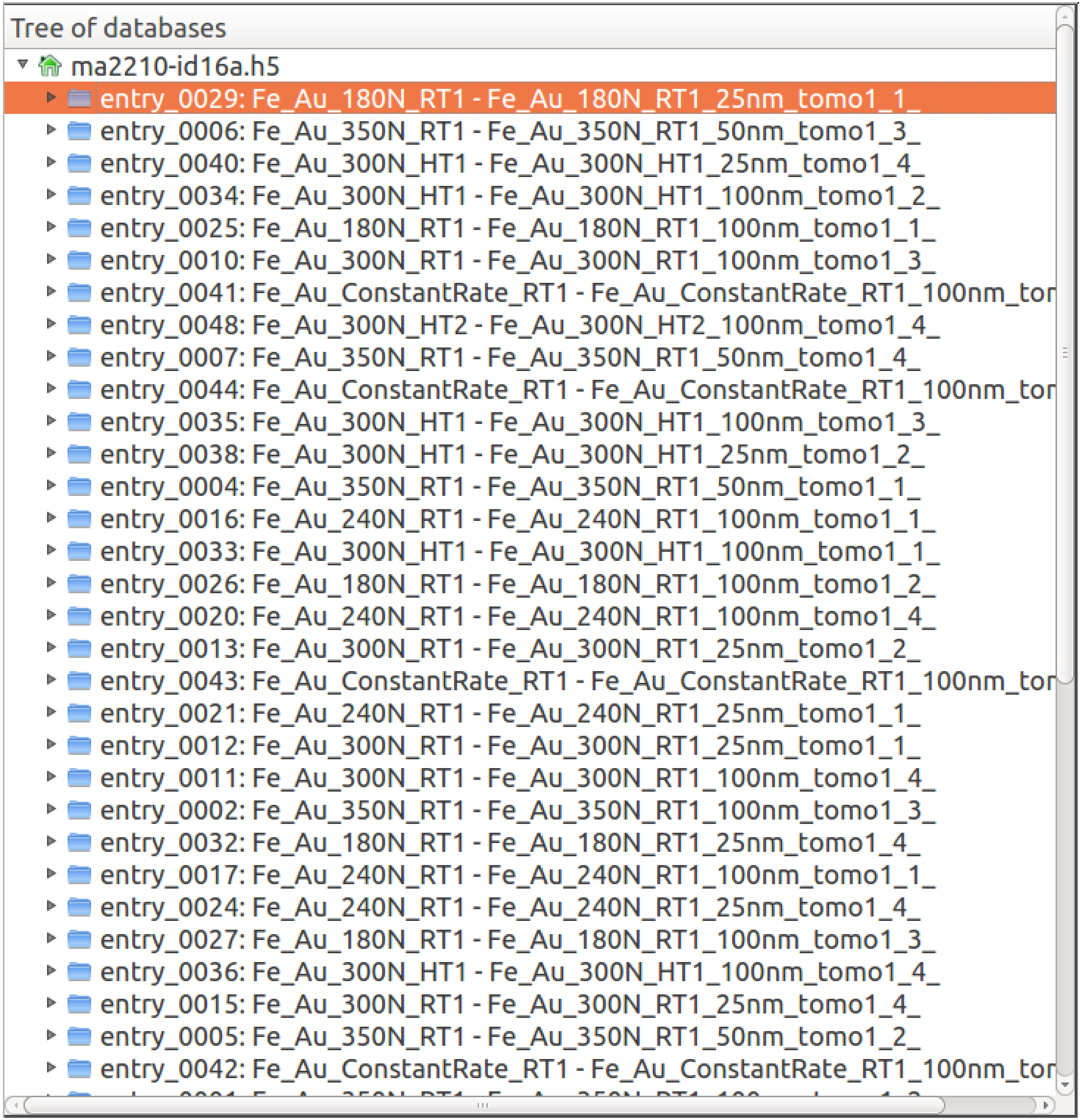
copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***



```
nano:20000/id16ni/metadata/tomo
${entry} = HDF5 file entry
title = Name of the scan
experiment_identifier = Proposal code
start_time = Scan starting date
end_time = Scan ending date
collection_time = [id16ni/spec/zaptomo/TOMO_EXPTIME]
sample [NXsample]
    name = Name of the sample
    distance = ${sourceSampleDistance} [Number, mm, NX_LENGTH]
    positioner [NXpositioner]
        name = ${sample_motors} [String]
        value = [id16ni/emotion_flexdc/srot/position, id16ni/emotion_pmd206_1/sx/position, id16r
    sensor [NXsensor]
        name = ${sample_sensors_labels} [String]
        value = [id16ni/hpzdrift/hpz_tz/DriftCorrection]
    pixel_size = [id16ni/ImagePixelSize/frelon1/pixelWidth]
    focus_position = ${sx0} [Number, mm]
    vacuum [NXenvironment]
        sensor [NXsensor]
            name = ${vacuum_labels} [String]
            value = [id16ni/v-pen/111/pressure]
instrument [NXinstrument]
    name = ID of the beamline
    detector [NXdetector]
        name = ${cameraName} [String]
        positioner [NXpositioner]
            name = ${detectors_motors} [String]
            value = [id16ni/motor/img1x/position, id16ni/motor/img1y/position, id16ni/motor/img1z
        source_distance = ${sourceDetectorDistance} [Number, mm]
    insertion_device [NXinsertion_device]
        gap = [orion:10000/id/id/16ni/U18-3C_GAP_Position, orion:10000/id/id/16ni/U18-3D_GAP_F
        name = ${insertionDeviceName} [String]
    source [NXsource]
        name = ESRF
        type = Synchrotron X-ray Source
        mode = [orion:10000/fe/id/16/SR_Filling_Mode]
        current_start = [orion:10000/fe/id/16/SR_Current]
        current_end = [orion:10000/fe/id/16/SR_Current]
    monochromator [NXmonochromator]
        name = [id16ni/energy/multilayer/positionId]
```

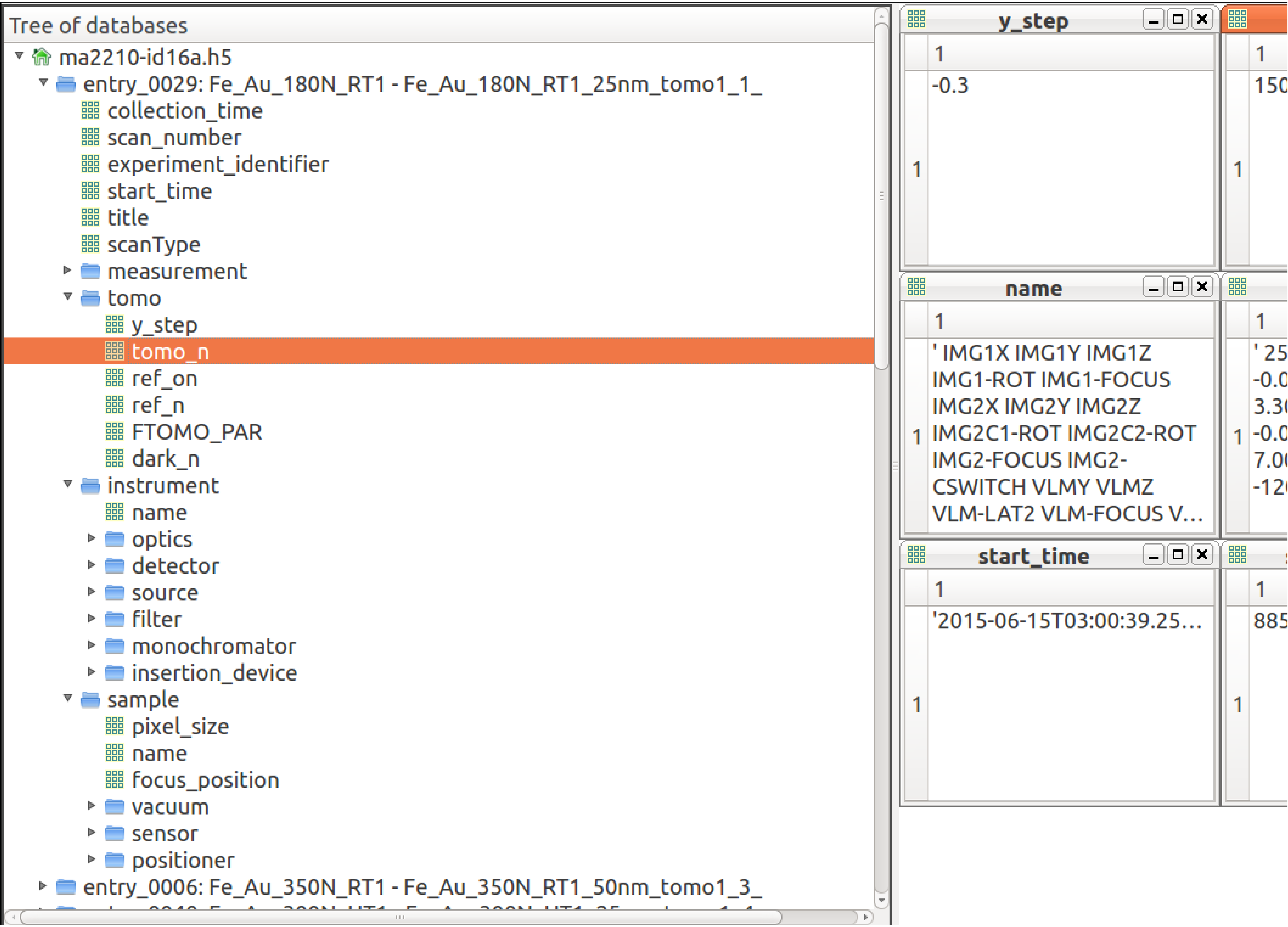
```
nano:20000
dserver
ID16NI
    adc
    attenuator
    backgroundsubstraction
    basler
    beamviewer
    bpm4q
    brake
    bsh
    bvmotors
    capa
    ctaccumulation
    ctacquisition
    ctbuffer
    ctconfig
    ctcontrol
    ctevent
    ctimage
    ctsaving
    ctshutter
    ctvideo
    detector
    devTCP
    diodet
    discretemotor
    elettra
    emotion
    emotion_cyril
    emotion_e517a
    emotion_e517b
    emotion_e753
    emotion_flexdc
    emotion_pmd206_1
    emotion_pmd206_2
    emotion_sp
    emotion_spz
    encoder
    energy
    experiment-itlk
    experiment-wago
    flatfield
    foil
```

ID16A example dataset



Tree of databases

```
▾ ma2210-id16a.h5
    ▸ entry_0029: Fe_Au_180N_RT1 - Fe_Au_180N_RT1_25nm_tomo1_1_
    ▸ entry_0006: Fe_Au_350N_RT1 - Fe_Au_350N_RT1_50nm_tomo1_3_
    ▸ entry_0040: Fe_Au_300N_HT1 - Fe_Au_300N_HT1_25nm_tomo1_4_
    ▸ entry_0034: Fe_Au_300N_HT1 - Fe_Au_300N_HT1_100nm_tomo1_2_
    ▸ entry_0025: Fe_Au_180N_RT1 - Fe_Au_180N_RT1_100nm_tomo1_1_
    ▸ entry_0010: Fe_Au_300N_RT1 - Fe_Au_300N_RT1_100nm_tomo1_3_
    ▸ entry_0041: Fe_Au_ConstantRate_RT1 - Fe_Au_ConstantRate_RT1_100nm_tor
    ▸ entry_0048: Fe_Au_300N_HT2 - Fe_Au_300N_HT2_100nm_tomo1_4_
    ▸ entry_0007: Fe_Au_350N_RT1 - Fe_Au_350N_RT1_50nm_tomo1_4_
    ▸ entry_0044: Fe_Au_ConstantRate_RT1 - Fe_Au_ConstantRate_RT1_100nm_tor
    ▸ entry_0035: Fe_Au_300N_HT1 - Fe_Au_300N_HT1_100nm_tomo1_3_
    ▸ entry_0038: Fe_Au_300N_HT1 - Fe_Au_300N_HT1_25nm_tomo1_2_
    ▸ entry_0004: Fe_Au_350N_RT1 - Fe_Au_350N_RT1_50nm_tomo1_1_
    ▸ entry_0016: Fe_Au_240N_RT1 - Fe_Au_240N_RT1_100nm_tomo1_1_
    ▸ entry_0033: Fe_Au_300N_HT1 - Fe_Au_300N_HT1_100nm_tomo1_1_
    ▸ entry_0026: Fe_Au_180N_RT1 - Fe_Au_180N_RT1_100nm_tomo1_2_
    ▸ entry_0020: Fe_Au_240N_RT1 - Fe_Au_240N_RT1_100nm_tomo1_4_
    ▸ entry_0013: Fe_Au_300N_RT1 - Fe_Au_300N_RT1_25nm_tomo1_2_
    ▸ entry_0043: Fe_Au_ConstantRate_RT1 - Fe_Au_ConstantRate_RT1_100nm_tor
    ▸ entry_0021: Fe_Au_240N_RT1 - Fe_Au_240N_RT1_25nm_tomo1_1_
    ▸ entry_0012: Fe_Au_300N_RT1 - Fe_Au_300N_RT1_25nm_tomo1_1_
    ▸ entry_0011: Fe_Au_300N_RT1 - Fe_Au_300N_RT1_100nm_tomo1_4_
    ▸ entry_0002: Fe_Au_350N_RT1 - Fe_Au_350N_RT1_100nm_tomo1_3_
    ▸ entry_0032: Fe_Au_180N_RT1 - Fe_Au_180N_RT1_25nm_tomo1_4_
    ▸ entry_0017: Fe_Au_240N_RT1 - Fe_Au_240N_RT1_100nm_tomo1_1_
    ▸ entry_0024: Fe_Au_240N_RT1 - Fe_Au_240N_RT1_25nm_tomo1_4_
    ▸ entry_0027: Fe_Au_180N_RT1 - Fe_Au_180N_RT1_100nm_tomo1_3_
    ▸ entry_0036: Fe_Au_300N_HT1 - Fe_Au_300N_HT1_100nm_tomo1_4_
    ▸ entry_0015: Fe_Au_300N_RT1 - Fe_Au_300N_RT1_25nm_tomo1_4_
    ▸ entry_0005: Fe_Au_350N_RT1 - Fe_Au_350N_RT1_50nm_tomo1_2_
    ▸ entry_0042: Fe_Au_ConstantRate_RT1 - Fe_Au_ConstantRate_RT1_100nm_tor
```



```
Tree of databases
▾ ma2210-id16a.h5
  ▾ entry_0029: Fe_Au_180N_RT1 - Fe_Au_180N_RT1_25nm_tomo1_1_
      collection_time
      scan_number
      experiment_identifier
      start_time
      title
      scanType
    ▸ measurement
    ▾ tomo
        y_step
        tomo_n
        ref_on
        ref_n
        FTOMO_PAR
        dark_n
    ▾ instrument
        name
      ▸ optics
      ▸ detector
      ▸ source
      ▸ filter
      ▸ monochromator
      ▸ insertion_device
    ▾ sample
        pixel_size
        name
        focus_position
      ▸ vacuum
      ▸ sensor
      ▸ positioner
  ▸ entry_0006: Fe_Au_350N_RT1 - Fe_Au_350N_RT1_50nm_tomo1_3_
```

```
y_step
     1
1    -0.3
```

```
     1
1    150
```

```
name
     1
1    ' IMG1X IMG1Y IMG1Z
     IMG1-ROT IMG1-FOCUS
     IMG2X IMG2Y IMG2Z
     IMG2C1-ROT IMG2C2-ROT
     IMG2-FOCUS IMG2-
     CSWITCH VLMY VLMZ
     VLM-LAT2 VLM-FOCUS V...
```

```
     1
1    ' 25
     -0.0
     3.30
     -0.0
     7.00
     -120
```

```
start_time
     1
1    '2015-06-15T03:00:39.25...
```

```
     1
1    885
```

Current status of Metadata

- Nexus/HDF5/icat system running on 2 beamlines with work started for 2 more

- Techniques implemented so far :
  - holo-tomography, nano-fluorescence, nano-diffraction, generic scanning

- Techniques to implement next :
  - ptychography, full-field diffraction, fluorescence, saxs, 3d x-ray diffraction, ...

Estimate of Metadata production @ ESRF

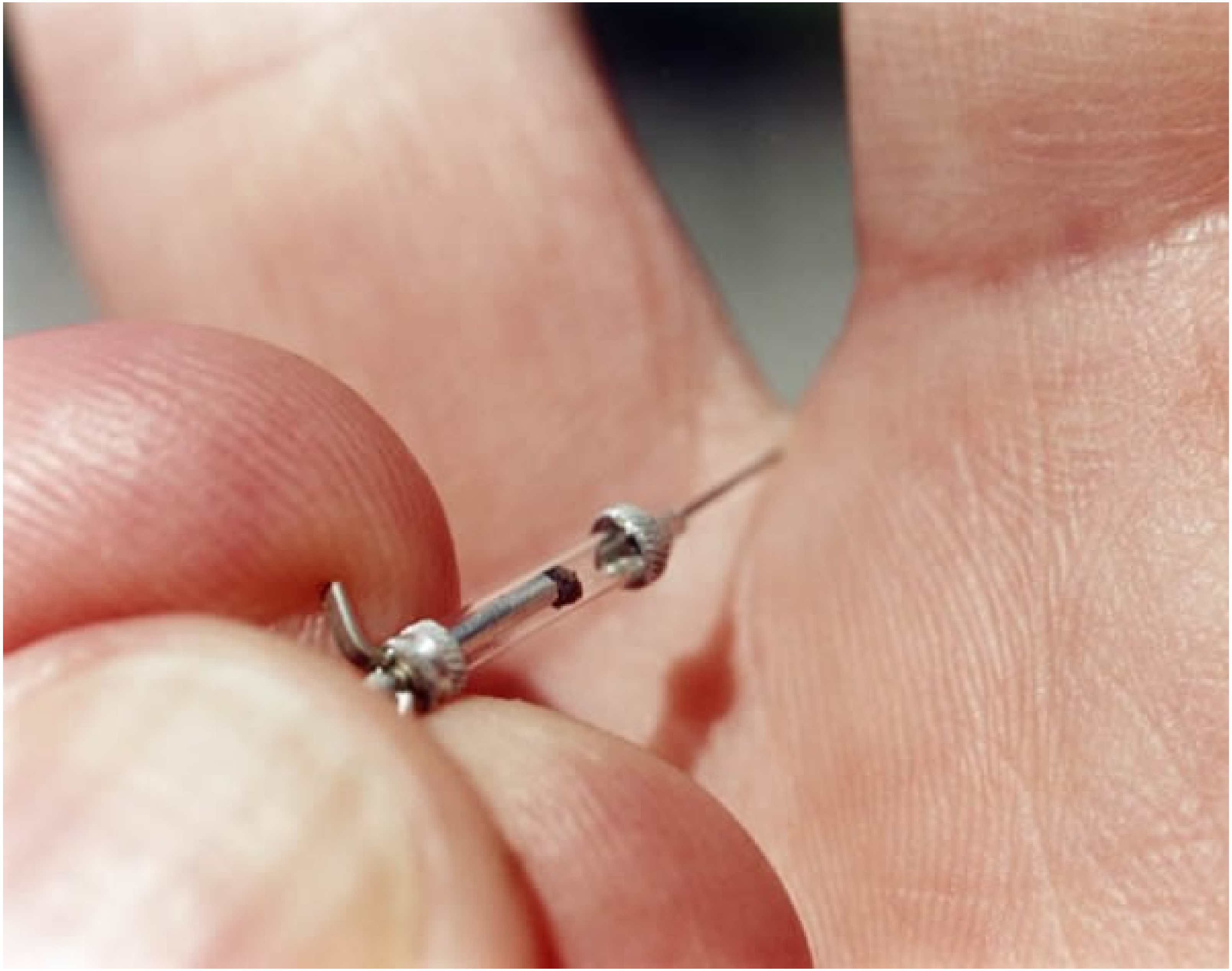
- Projection based on 1 beamline
- Database size after 10 years < 2 TB

		As of 2015/02/11	Per week of operation per beamline	Per year of operation per beamline	Per week of operation on all beamline	Per year of operation on all beamlines
Proposal		19	1.5	57	58.5	2280
Users		69	5.3	207	212.3	8280
Sample		139	10.7	417	427	16680
Dataset		1917	147	5751	5898.5	230040
Datafile		1807096	139007	5421288	5560295	216851520
Parameter		55172	4244.0	165516	169760.0	6620640
Database size	GB	1	.1	3	3.1	120

Storing Metadata

- Metadata will be stored in a **single master file** (HDF5) per experiment + in **icat database**
- **Data analysis** will be able to **access** the **raw data** through the master file **via links**
- It will be possible to **regenerate** the **master file from icat** using the latest configuration file
- Storing metadata is **low cost** in terms of **disk storage** but needs **human resources** to be maintained

Data and metadata policy



copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

Metadata policy

- One of hurdles to defining a data policy – **cost of storage** – is **NOT** an issue for Metadata
- **Metadata** needs to be **complete** so that data can be **re-analysed**
- **ESRF** is **committed** to defining and implenting a **metadata** and **data policy** during **Phase II**
- What **metadata policy** to apply when publishing metadata under a **DOI** ?

Metadata is key to progress

copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

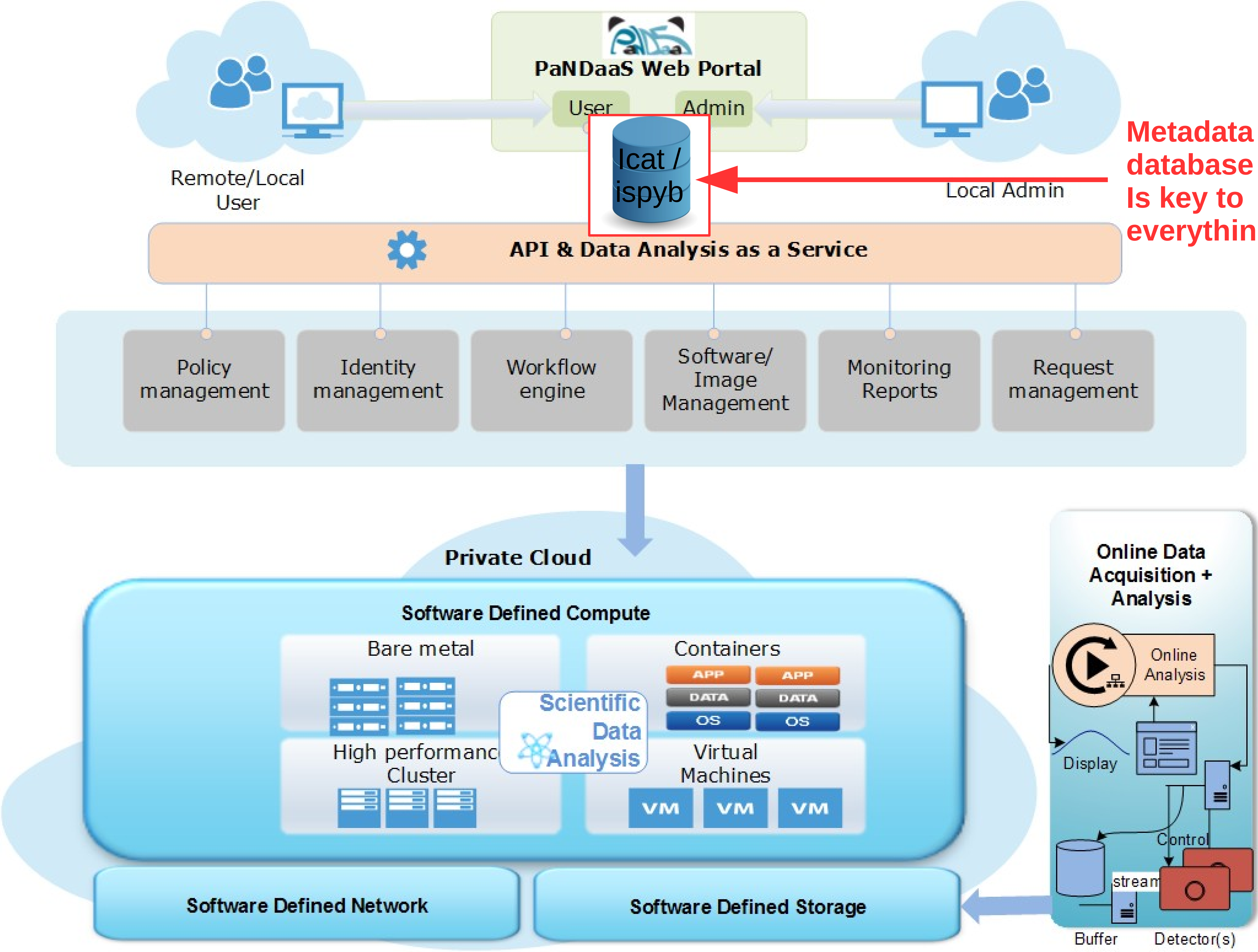
Metadata key to progress

- Any **new data based service** requires quality metadata :
  - Online data analysis
  - Automated workflows
  - Archiving + retrieval
  - Metadata mining
  - Re-analysis of data
  - Cloud-based services
  - Linking raw data to publications

Data Analysis as a Service

- Data is becoming increasingly difficult to take home
- Users are increasingly facing storage and performance issues
- New users have issues with software
- Solution = **leave data at the source and provide remote access to data and software**
- Sound familiar ? Cloud ... PaNDaaS* proposal

*project progress will be posted on http://pan-data.eu/



PaNDaaS Web Portal

User

Admin

Icat / ispyb

Remote/Local User

Local Admin

API & Data Analysis as a Service

Policy management

Identity management

Workflow engine

Software/ Image Management

Monitoring Reports

Request management

Private Cloud

Software Defined Compute

Bare metal

Containers

High performance Cluster

Scientific Data Analysis

Virtual Machines

Software Defined Network

Software Defined Storage

Online Data Acquisition + Analysis

Metadata database Is key to everything

How this workshop can help

copyright **Cynthia Greig** (http://cynthiagreig.com) - ***Life Size***

How this workshop can help

- **Sharing** of **metadata schemes** for data analysis

- Define an international **Metadata policy**

- **Insist** on the need to curate **raw data**

- **Coordinate metadata** for new techniques related to **crystallography**

Links to other intiatives

- Similar efforts on going at :
  - **DESY**
  - **NSLS II**
  - others ...

Conclusion

- A **beamline specific approach** is **not enough**
- **Global site-wide approach** linked to a **database** is needed
- **Metadata** definitions need to be **flexible** and support **multi-technique** experiments
- Metadata is **more** than **image headers** – the answer is a **master file** with all metadata required for data analysis
- Keeping **definitions coherent** is a **challenge**
- **Metadata policy** is unavoidable in the future

Credits

- **ESRF** - Christophe Cleva, Armando Solé, Peter Cloetens, Christian Nemoz, Cyril Guilloud, Roberto Homs, Olof Svensson, Jerome Kieffer, Peter Boesecke, Julio Cesar Da Silva, Alessandro Mirone

- Photographs - **Cynthia Greig**