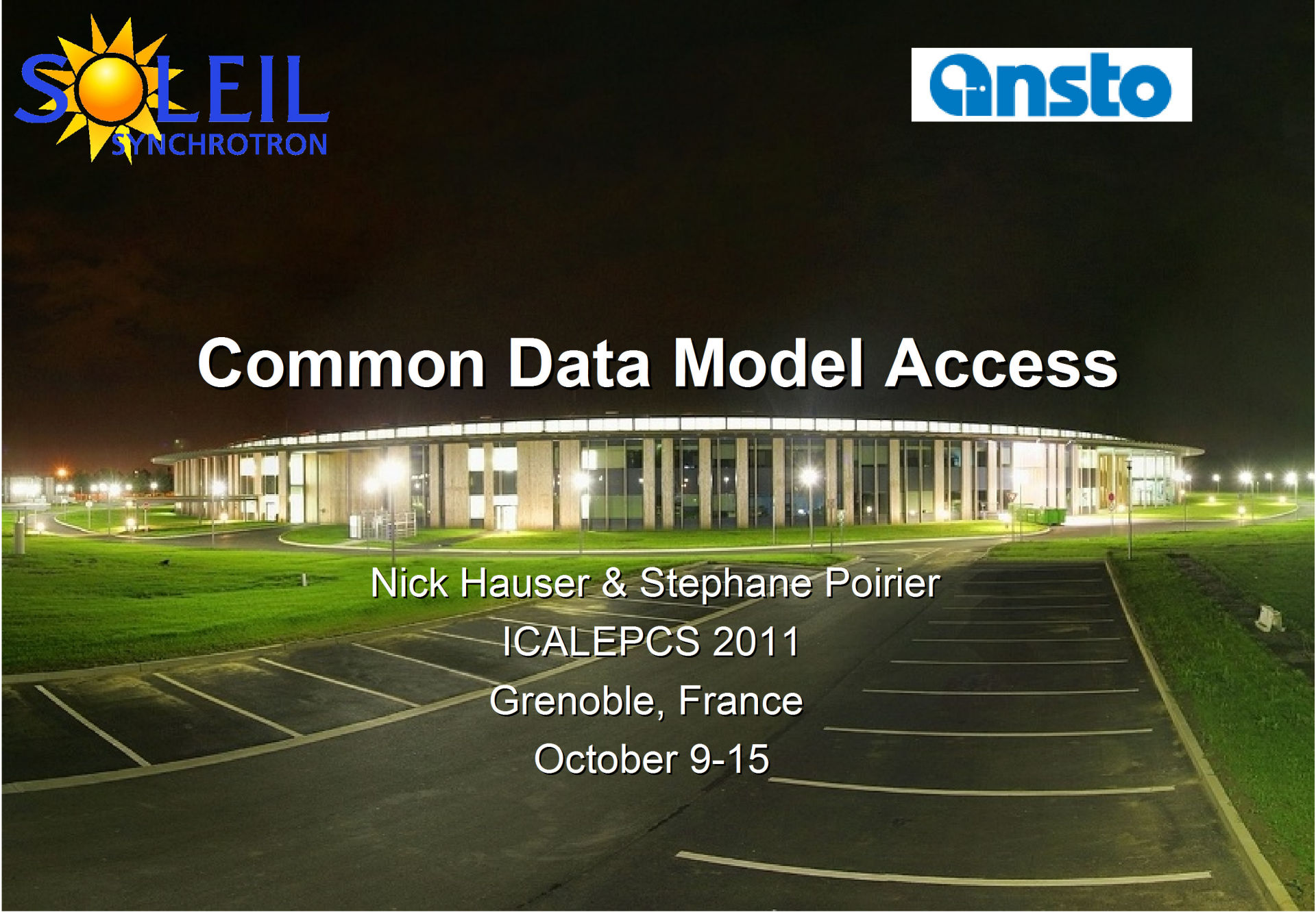


Common Data Model Access

Nick Hauser & Stephane Poirier

ICALEPCS 2011

Grenoble, France

October 9-15

Where is ANSTO?
Approx 16,500km from here

SOLEIL SYNCHROTRON

ansto
Nuclear-based science benefiting all Australians

What is ANSTO?
Australian Nuclear Science & Technology

There is nothing more important than data structures

I will, in fact, claim that the difference between a ***bad programmer*** and a good one is whether he considers his code or his data structures more important. Bad programmers worry about the code. Good programmers worry about data structures and their relationships.

Torvalds, Linus (2006-06-27). Message to Git mailing list.

ICALEPCS 2011
Collaborate, discover, enjoy.

France and Australia collaborate in a common data model.

Across 9 time zones, and 16,968km

Maybe they will collaborate in the Rugby World Cup final?

Collaboration and competition. Clearly both are important.

We need a gene pool.

Genesis. In the beginning…

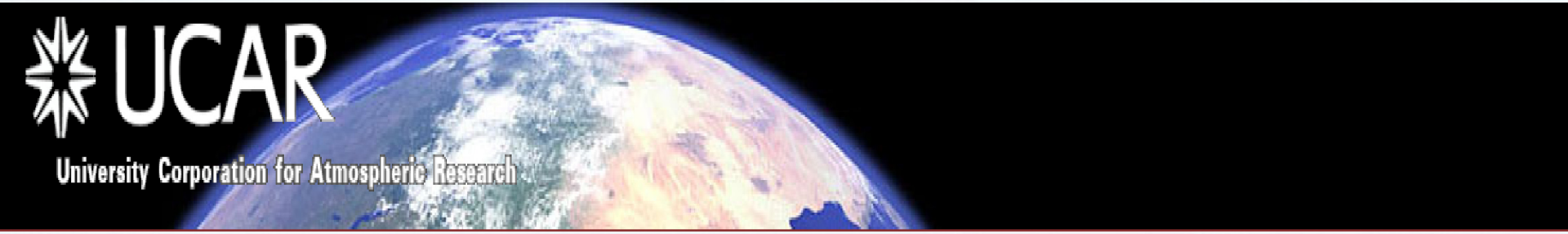
Give me a way to read my experimental data into any analysis package!

Not a new idea

just an implementation of an old one. Yawn.

Can you wake me up when this talk is over?

Climate scientists have been doing this for years.



unidata
providing data services, tools and cyberinfrastructure leadership

netCDF & OGC

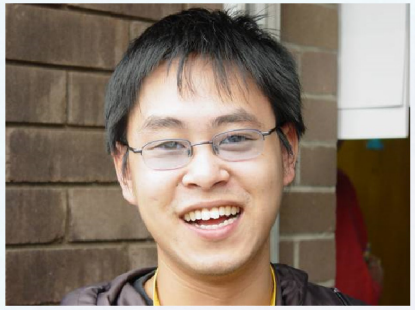
We borrowed their ideas... in particular, the interfaces, and extended them into our domain.

SOLEIL SYNCHROTRON

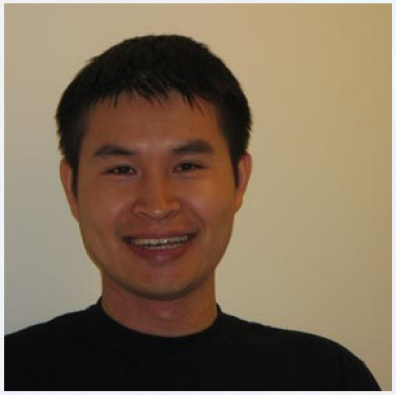
ansto
Nuclear-based science benefiting all Australians

The fable of the two developers

Once upon a time, there were 2 developers...

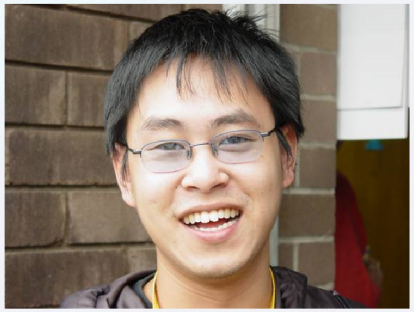


Data plug-in developer

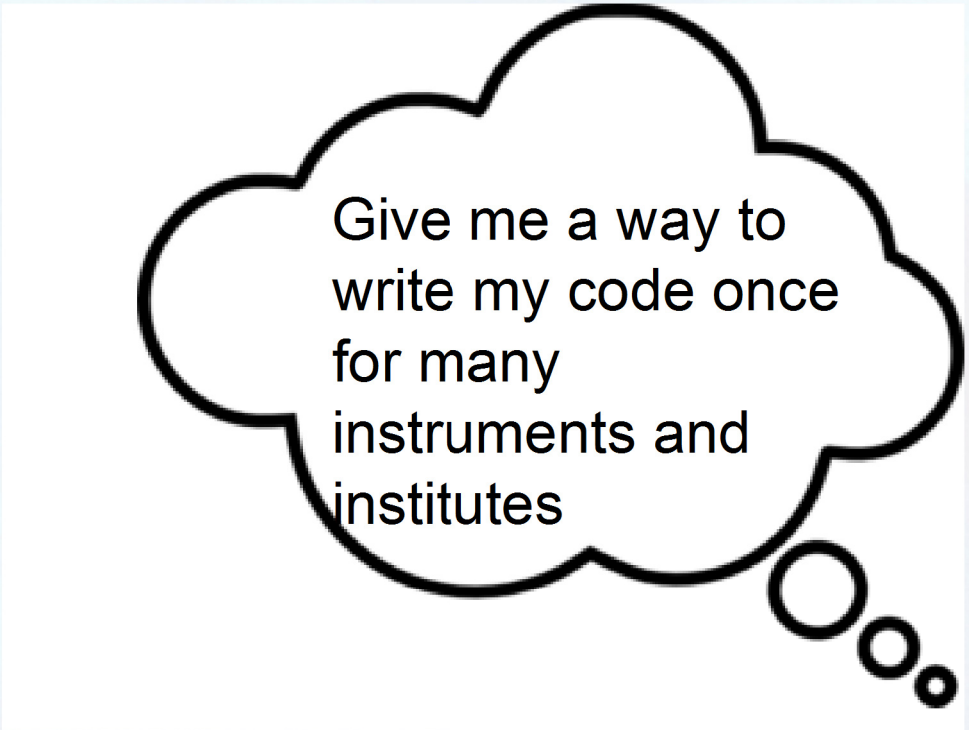


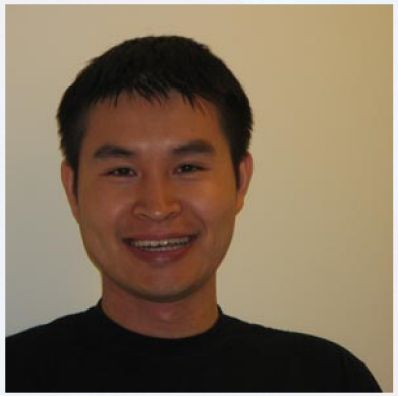
Data reduction developer

Give me an API to read my institute's data into an object-oriented data object



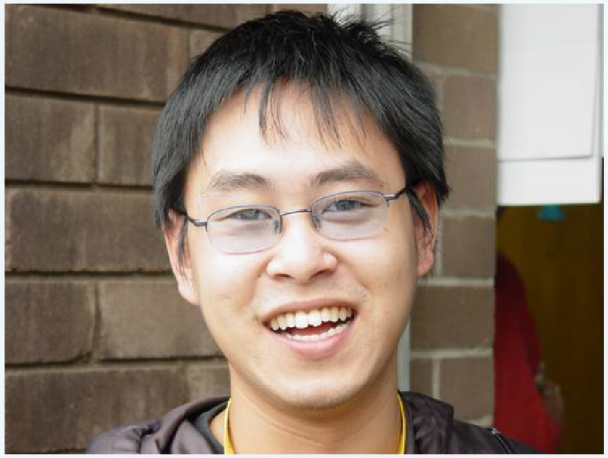
Data plug-in developer





Data reduction developer

ICALEPCS 2009.
The collaboration begins



Tony Lam, ANSTO meets…

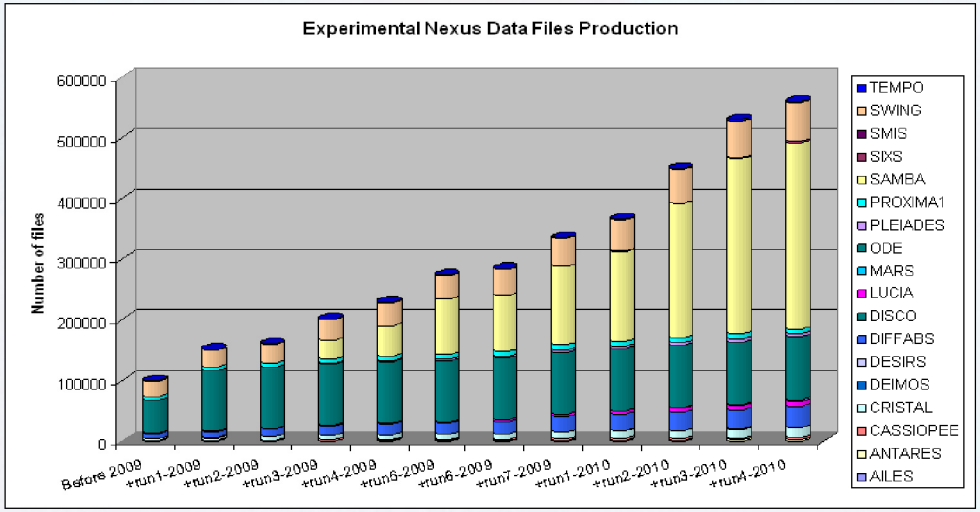
Majïd Ounsy, Soleil

The SOLEIL experience

NeXus/HDF5 data format is the “SOLEIL standard” on all our beamlines since the beginning

We defined a standard internal data file structure for experimental data storage

On acquisition side
all works fine !

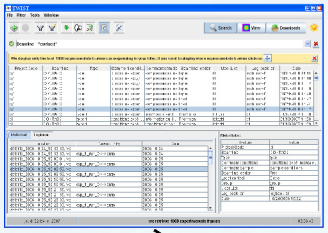


SOLEIL SYNCHROTRON

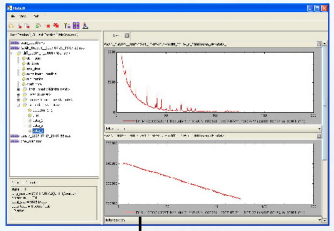
ansto
Nuclear-based science benefiting all Australians

How to exchange data?

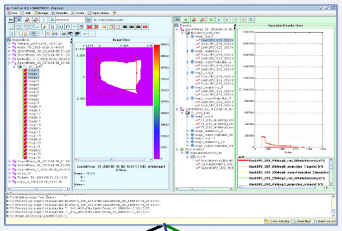
File retrieval



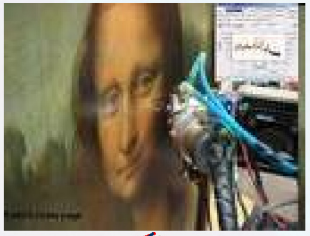
File browsing



SAXS Data Analysis foxtrot Application

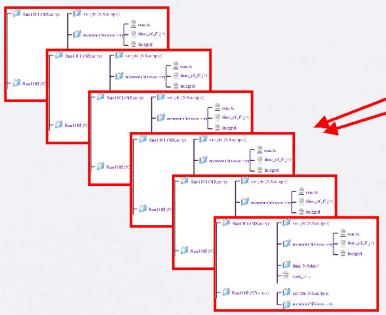


Data Analysis Application B

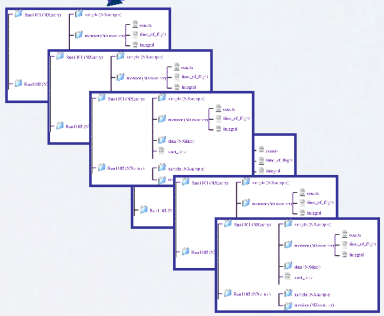


Data Analysis Application C

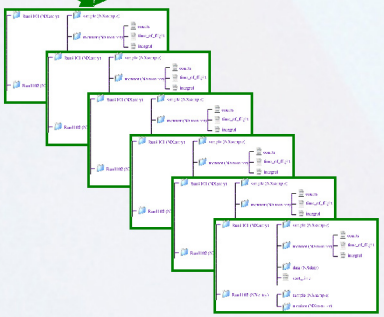
NeXus Application Interface



SOLEIL NeXus Files



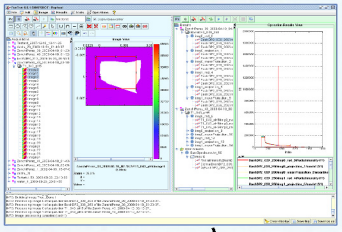
ESRF Files



DESY Files

Nuclear-based science benefiting all Australians

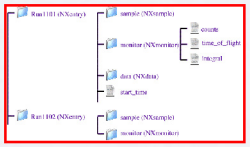
How to process data?



SAXS Data Analysis
foxtrot Application

NeXus Application Interface

Soleil NeXus file created on CRISTAL



Soleil NeXus file created on SWING

In reality we were blocked even before exchanging data with other institutes!!

On SWING beamline, data organisation had been fixed by a « SAXS oriented » data acquisition sequence

On CRISTAL beamline, data organisation inside the NeXus file had been fixed by a «Powder diffraction oriented » data acquisition sequence

SOLEIL SYNCHROTRON

ansto
Nuclear-based science benefiting all Australians

What is the CDMA?

A plug-in system that allows support of various data sources

An abstract interface for navigation through data sets

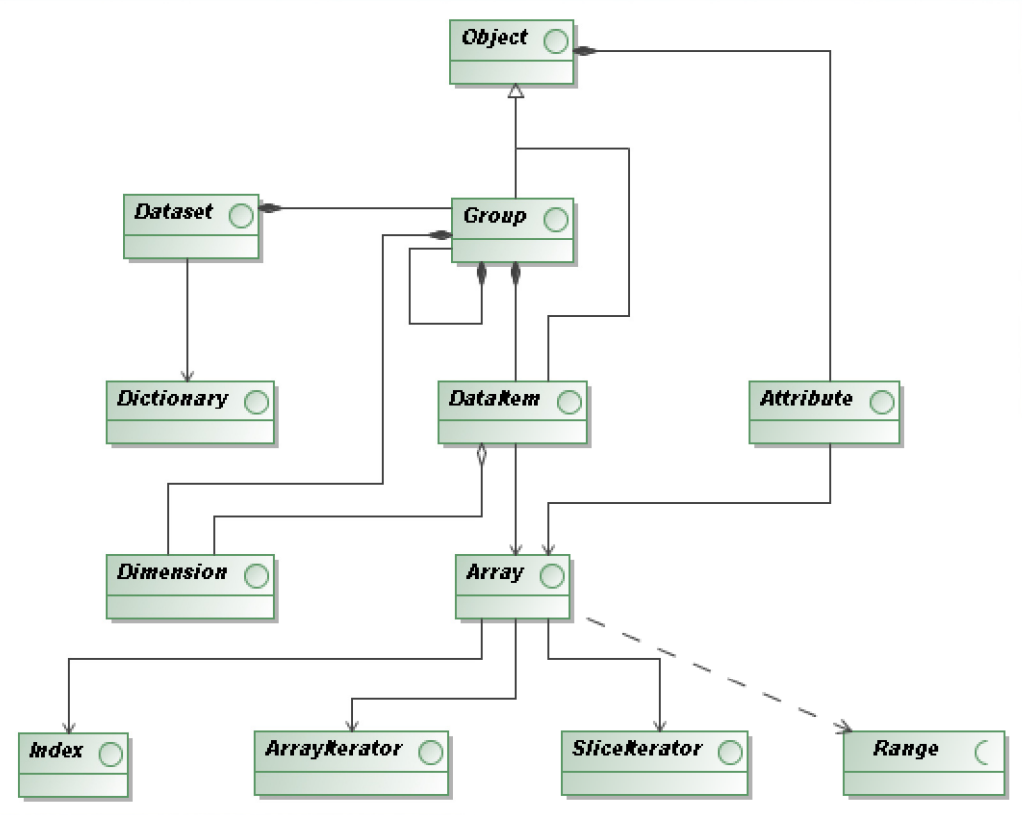
- Concrete classes are provided by data source plug-ins

A dictionary mechanism to retrieve measurements and technical values independently of the data structures

A manipulation class for array slicing and dicing

An utility class for array mathematics operations with error propagation calculation

The model - simple



One data model

2 API's

Reduction developer

Plug-in developer

Separation of concern

Life is easy

Plug-in system

A CDMA plug-in is a piece of code that adds the support of a specific data file format (e.g. NeXus, netCDF, Spec,...)

Plug-ins are loaded at run-time

A client application can load simultanously two (or more) files that required different plug-ins

Available plug-ins : HDF5 (support of netCDF & NeXus), EDF (ESRF Data Format, partially)

Dictionary mechanism

The main point of CDMA is to allow a data analysis application that does not care about file format.

We think it's not sufficient. Applications developers don't have to care about data structures.

To achieve this the CDMA API introduce the notion of ***dictionary***

A dictionary is

- A list of ***keywords*** (organized or not in separate groups, it doesn't matter)
- A set of association between ***keywords*** and ***data items*** in a specific data structure (NeXuS, NetCDF,...)

Dictionary mechanism

Dictionary documents are XML files

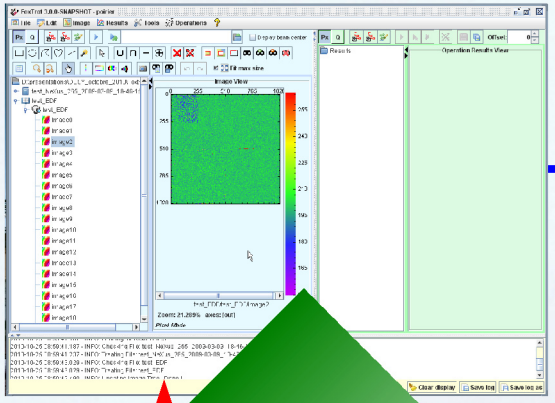
A dictionary is defined by the association of two documents:

- A first where some keywords are declared
  - It can be a flat list of keywords
  - It can be organized in a hierarchical way (a tree of keywords)
- A second where these keywords match scientific measurements paths in the data files
  - It's a map where keywords are linked to data structure

Each institute that contribute to this project have to provide:

- A plug-in for its own data format
- One or several mapping documents according a commonly accepted list of keywords

Dictionary mechanism



use

keywords Declaration file

```
<data-def name="Experiment type"> <!--
ex: EXAFS, SAXS,... -->

  <item key="raw_data"/>
  <item key="energy"/>

</data-def>
```

needs

CDMA

File format plugin

use

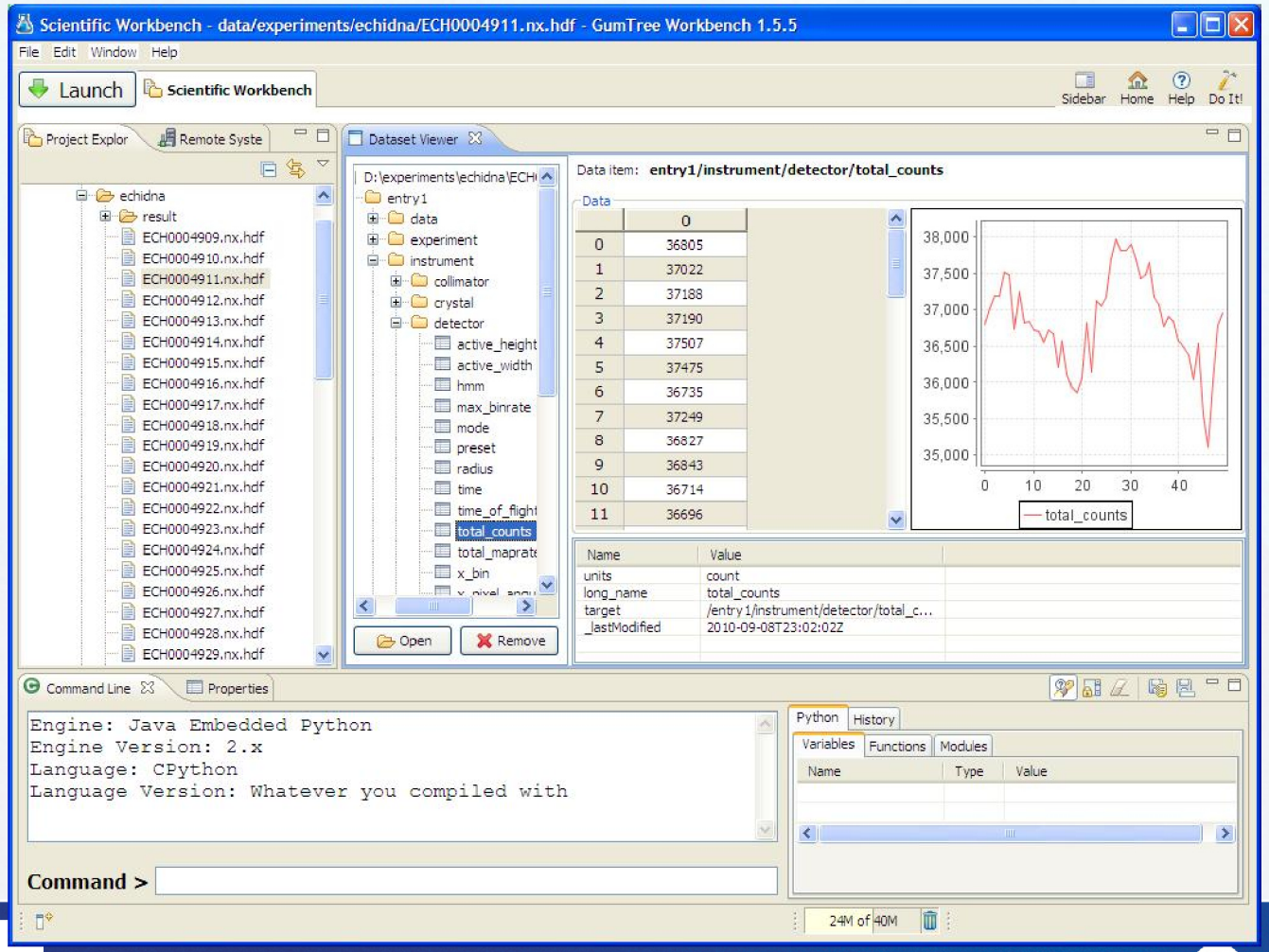
keywords Mapping file

```
<map-def name="Experiment type">  <!--
ex: EXAFS, SAXS,... -->

  <item key="raw_data">
    <path>path/to/raw_data</path>
  </item>

  <item key="energy">
    <path>path/to/wavelength</path>
    <call>WavelenghToEnergy</call>
  </item>

</map-def>
```

Dataset

0100110
1001110
0100110
11110...

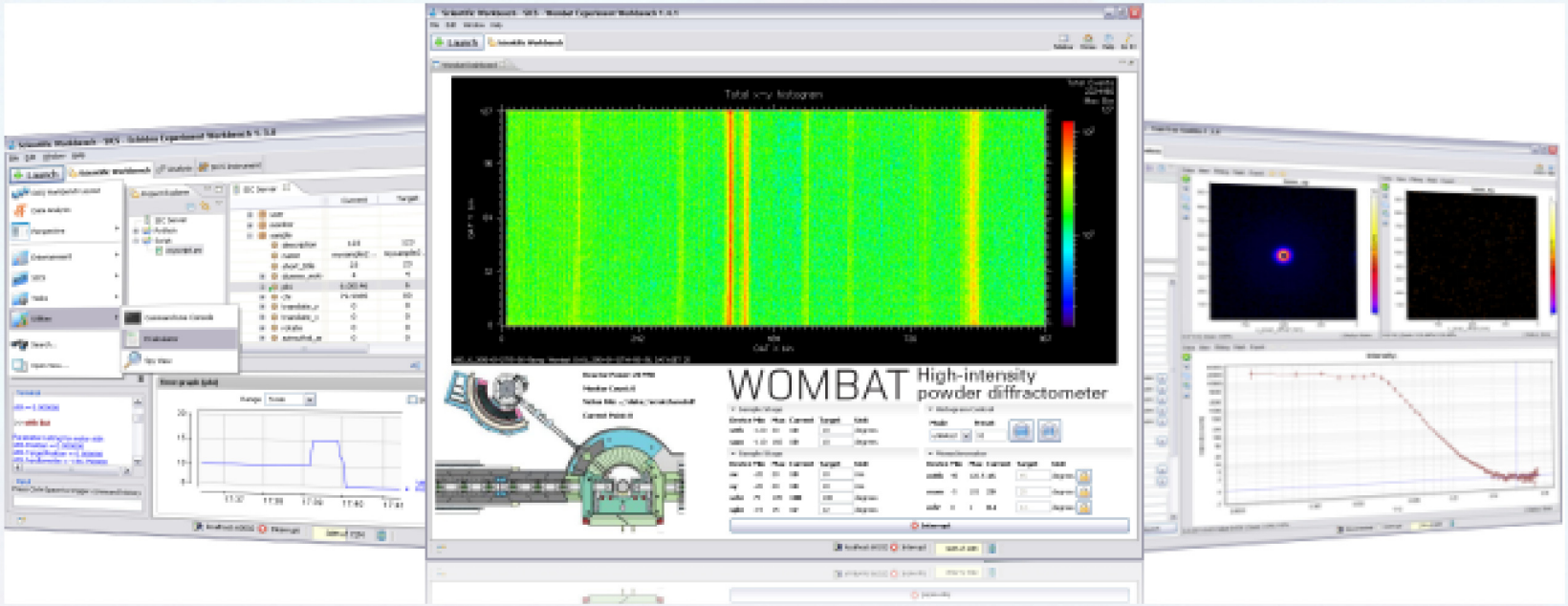
NeXus browse and visualisation



SOLEIL SYNCHROTRON

ansto
Nuclear-based science benefiting all Australians

Common data model as an OSGI bundle in the scientific workbench, Gumtree. In production since 2006



ICALEPCS 2011
Collaborate, discover, enjoy.

Software collaboration benefits from "**many eyes**".

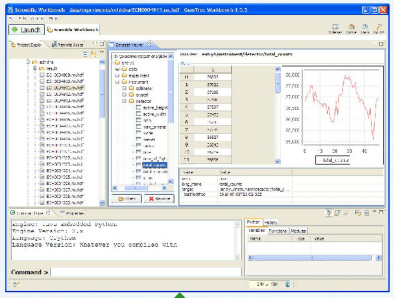
True for CDMA.

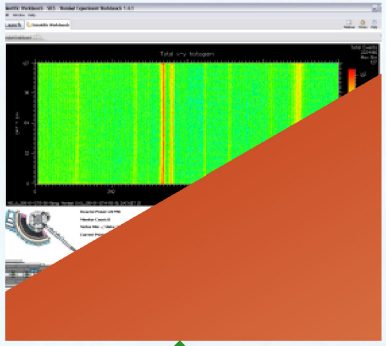
Databrowser

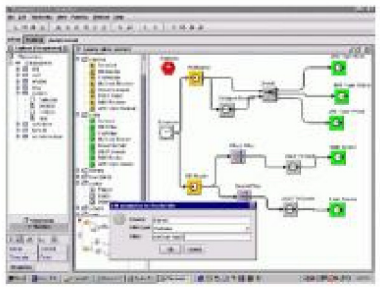
GumTree

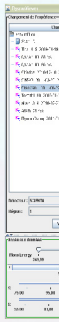
Passerelle

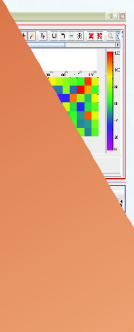
Data reduction











?

CDMA

TANGO

EPICS

?

The HDF Group

MySQL

?

Control systems

Data storage

SOLEIL SYNCHROTRON

ansto

Nuclear-based science benefiting all Australians

Conclusion

- ✓ CDMA in production at SOLEIL and ANSTO
- ✓ Sharing of data independent of data structure specification by NeXus or other standards body
- ✓ C++ port development in progress
- ✓ DESY will officially join the project giving resources
- ✓ Python and Matlab port to follow the C++ implementation
- ✓ Find us at “Common Data Model” Google Group

Common Data Model Access

Credits

ANSTO team - Paul Hathaway, Darren Kelly, Tony Lam, Norman Xiong

SOLEIL team - Stephane Poirier, Majid Ounsy, Clement Rodriguez, Alain Buteau