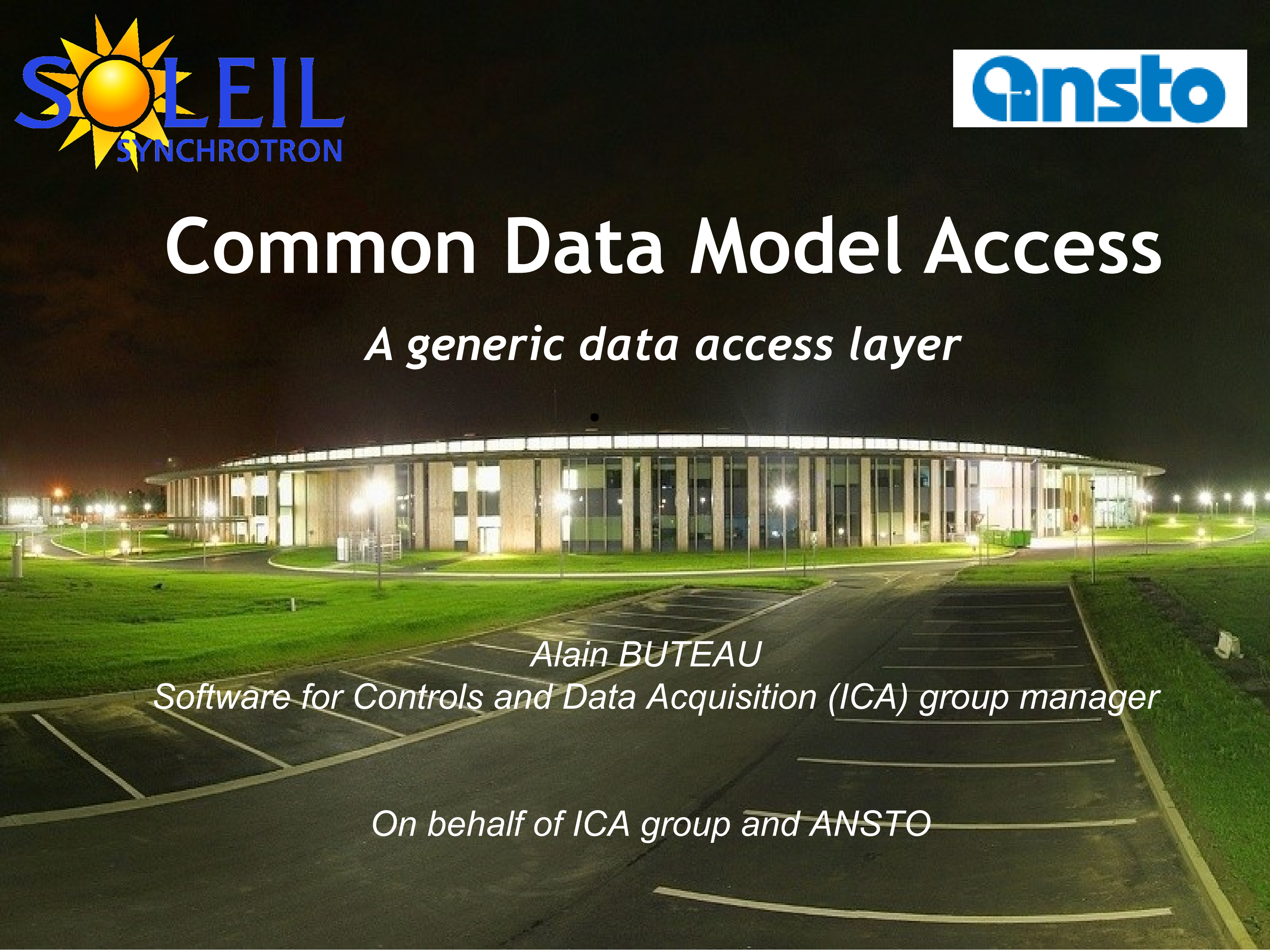


Common Data Model Access

A generic data access layer

Alain BUTEAU

Software for Controls and Data Acquisition (ICA) group manager

On behalf of ICA group and ANSTO

Agenda

- Motivations : Which problem do we need to solve
- Overview of the software architecture and the main concepts
- CommonDataModelAccess project management facts
- Conclusion
  - ➔Next technical steps
  - ➔How CDMA project can be actively integrated within PANDATA-ODI or other European project ?

Which problems do we need to solve ?

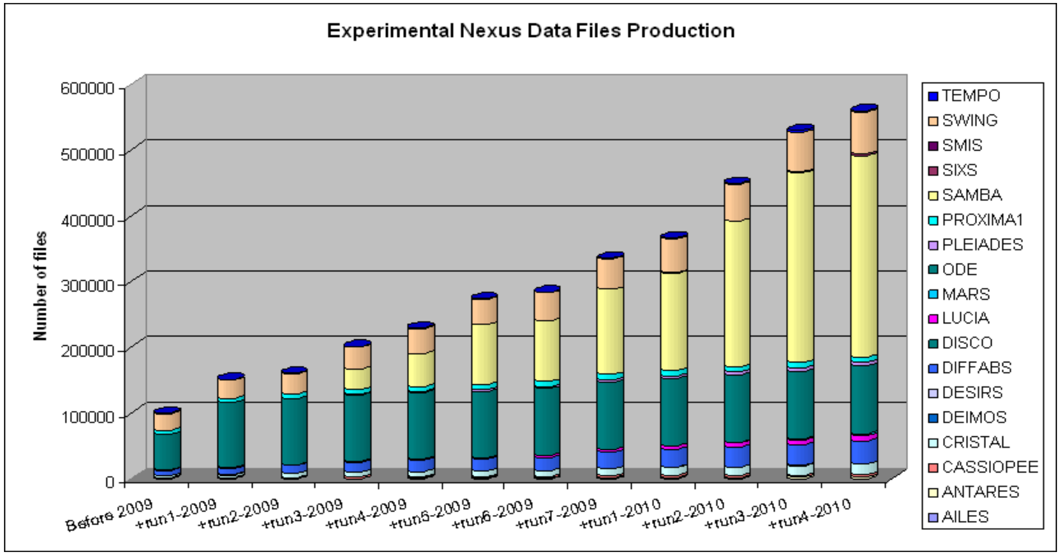
A solution for which problem ?

- Find solutions to data format issues from the **data analysis point of view**

- With two objectives in mind
  - ✓Find the most suitable ways to **exchange data** between our institutes
  - ✓Find the most suitable ways to **exchange reduction/analysis applications** between our institutes

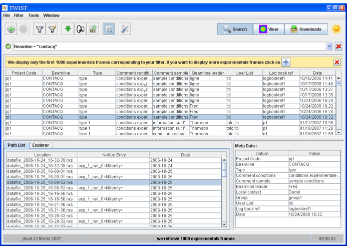
The foreseen solutions are :

- ✓Choose NeXus/HDF5 data format as the “SOLEIL standard” on all our beamlines
- ✓Define a standard internal data file structure for experimental data storage
- ✓**That 's the way we followed at SOLEIL during the last 7 years**

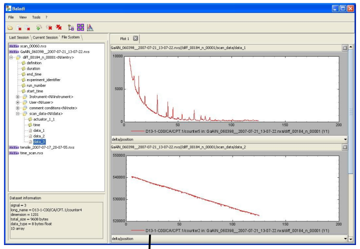


NeXus Files choice : Are we happy ?

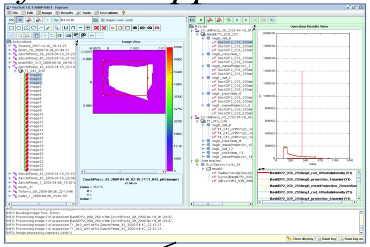
File retrieval



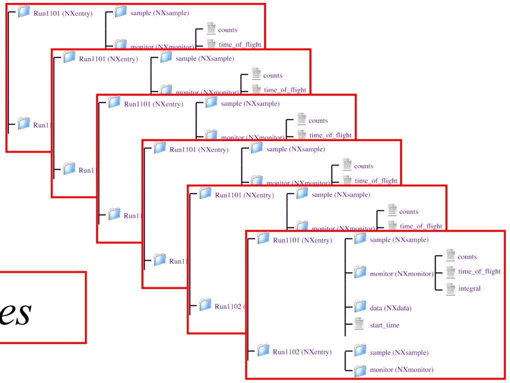
File browsing



SAXS Data Analysis foxtrot Application



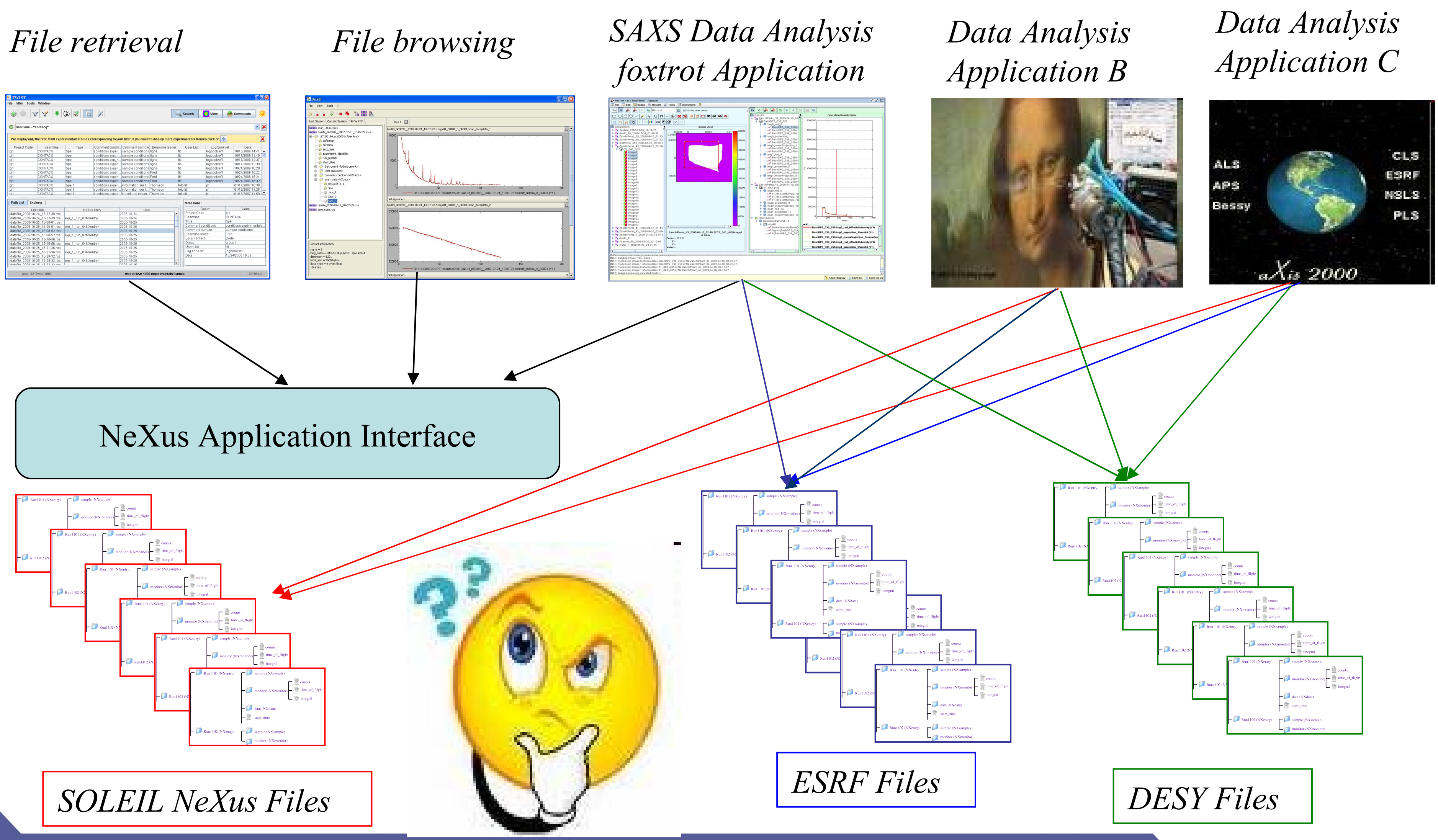
NeXus Application Interface



SOLEIL NeXus Files

- NeXus is a good and efficient storage format
- Thanks to a unique API and a « SOLEIL standardized internal data organization », we could :
  - ✓ *develop common software solutions*
  - ✓ *Decouple the development of Acquisition softwares from Data Analysis software*

NeXus Files choice : Are we happy ?



File retrieval

File browsing

SAXS Data Analysis foxtrot Application

Data Analysis Application B

Data Analysis Application C

NeXus Application Interface

SOLEIL NeXus Files

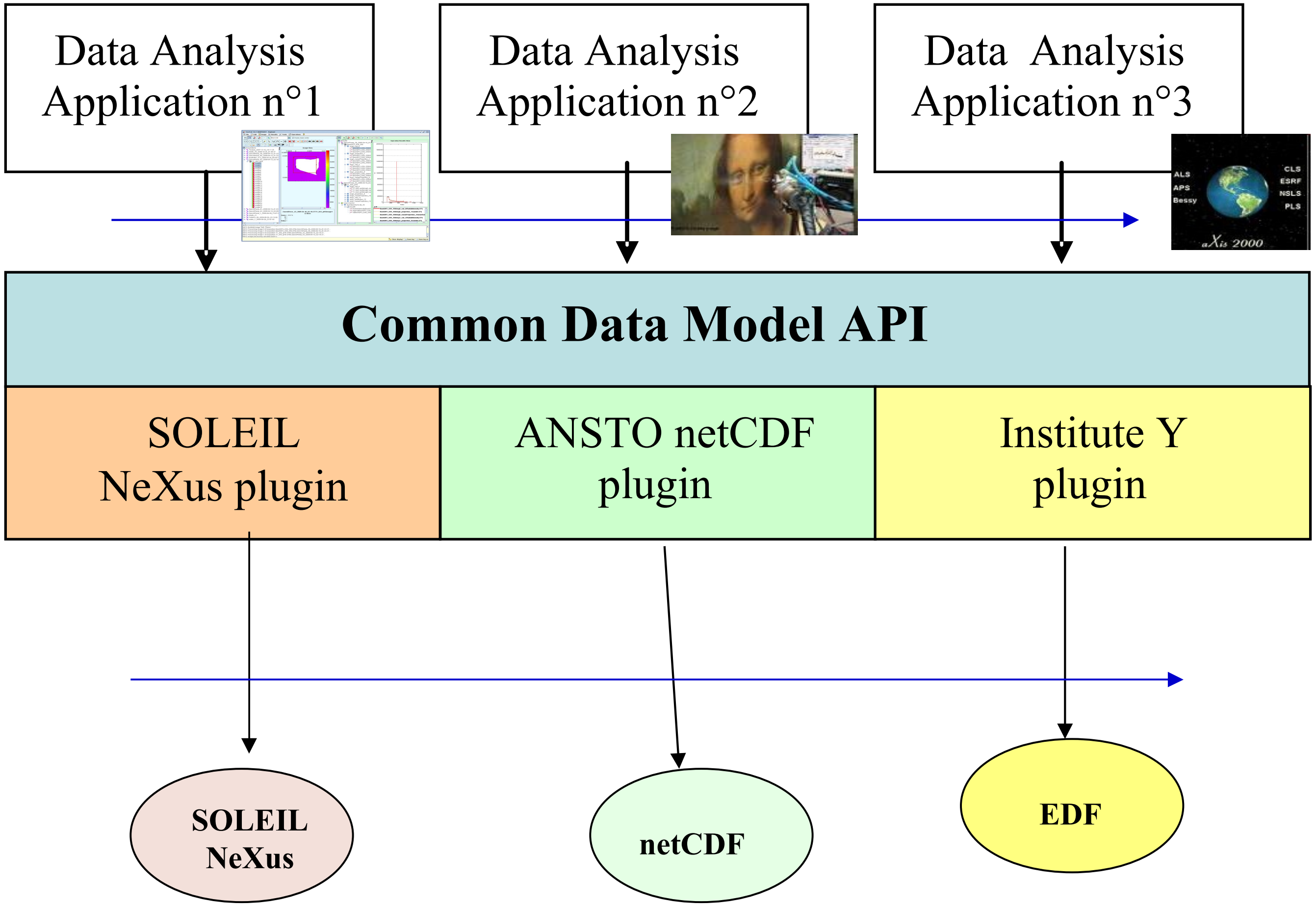
ESRF Files

DESY Files

Our proposal :
Introduce an indirection layer

Application developper adapt his application to the Common DataAccess API ONLY ONCE

Each institute implement plugin for each of its own data formats



Data Analysis Application n°1

Data Analysis Application n°2

Data Analysis Application n°3

Common Data Model API

SOLEIL NeXus plugin

ANSTO netCDF plugin

Institute Y plugin

SOLEIL NeXus

netCDF

EDF

CDMA concepts

CDMA hides the physical container of data

- A plug-in system that allows support of various data file formats (HDF5, CBF, EDF, etc ..)
- An abstract interface for navigation through data sets
  - ➔Concrete classes are provided by data source plug-ins

Introducing the dictionary mechanism

- The main point of CDMA is to allow a data analysis application to not care about physical file format.
- We think it's not sufficient. Developers of applications shouldn't have to care about data structures.
- To achieve this, the CDMA API introduces the notion of *dictionary*
- A dictionary is
  - Some ***keywords***
  - A set of associations between those ***keywords*** and ***data paths*** for a specific data structure (NeXuS, NetCDF,...)
  - Please see a keyword as a named scientific concept.

The dictionary mechanism

- Dictionary are XML files
- A dictionary is defined by the association of two files:
  - ➔a file where some keywords are declared
    - ▶ It can be organized in a hierarchical way (a tree of keywords)
    - ▶ It can be a flat list of keywords
  - ➔a file where these keywords match scientific measurements paths in the data files
    - ▶ It's a *map* where keywords are linked to data structure

Keywords declaration file

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


    <item key="user-name"/>
    <item key="e-mail"/>


    <item key="facility-name"/>
    <item key="facility-type"/>         <!-- 'X-ray', 'Neutron' -->


    <item key="energy"/>

    <item key="raw-data"/>


</data-def>
```

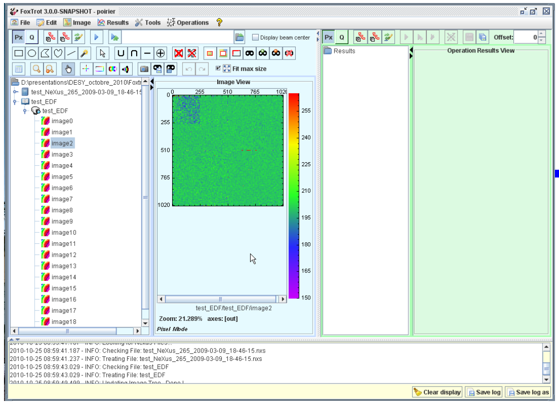
- This document defining a list of keywords (node 'data-def') that could be generic to a scientific domain and should be distributed to application developers. It has to be discussed by the community.

Keywords mapping

```
<map-def name="Experiment name">            <!-- ex: EXAFS, SAXS,... -->
  <item key ="user-name">
    <path>/path/to/user/name</path>
  </item>
  <item key ="energy">
    <path>/path/to/monochromator/wavelength</path>
    <call>WavelengthToEnergy</call>         <!-- plugin method -->
  </item>
  <item key="rawdata">
    <path>path/to/the/rawdata</path>
  </item>
  ...
</map-def>
```

- The second document is related to a particular file format/plug-in
- It's the responsibility of institutes producing data to provide this mapping according to a commonly accepted set of keywords

Dictionary schema



```
2010-10-25 08:59:41.187 - INFO: Checking File: test_NeXus_265_2009-03-09_18-46-15.nxs
2010-10-25 08:59:41.237 - INFO: Treating File: test_NeXus_265_2009-03-09_18-46-15.nxs
2010-10-25 08:59:43.029 - INFO: Checking File: test_EDF
2010-10-25 08:59:43.029 - INFO: Treating File: test_EDF
```

keywords Declaration file

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

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

</data-def>
```

CDMA

Files format plugin

0100110
1001110
0100110
11110...

keywords Mapping file

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

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

</map-def>
```

Project Management facts

Project history

- Collaboration between SOLEIL & ANSTO
  - → Started at the end of 2009
  - → When ANSTO sent to SOLEIL its current implementation of the GumTree Data Model.
- Since then 2 major versions of the java implementation of CDMA
  - →They are used in operation at SOLEIL and ANSTO
  - →Documentation is available (architecture and concepts , and a plugin developers guide)
- SOLEIL started the C++ port 4 months ago

Development plan for C++ implementation

- The C++ implementation is not complete yet
- Are missing :
  - → A complete dictionary implementation
  - →To complete the NeXus plug-in
- End of Q1-2012 seems to be a reasonable date to have a V1.0
- Then 2 tasks are to be done (it maybe done by other interested institutes !!)
  - →Python port
  - →Matlab port

Development plan on the Java side

- Enhancements to the current java implementation are foreseen
  - ➔ Roadmap will be discussed on Google group
- On the client side, the DataBrowser developed by ANSTO will be enhanced
  - ➔for example to take advantage of navigating through datasets using the dictionnary mechanism
- Many existing java applications/frameworks are candidates to use CDMA as one of their data source layer
  - ➔GDA , DAWB, Passerelle workflow engine,etc ..
  - ➔Enriching the CDMA ecosystem for newcomers who can then use these applications with a small development cost (the plugin)

Source code sharing

- Today the source code is hosted on the CodeHaus public repository
- It is for now a sub-project of GumTree.
  - ➔The SVN repository is localized here: https://svn.codehaus.org/gumtree/datamodel/
- Now that project is mature enough SOLEIL and ANSTO decided to move as a separate SVN project
  - ➔The move is in progress

Organisation of source code co-development

- CDMA core java : ANSTO is in charge of releases
- CDMA core C++ : SOLEIL is charge of releases
- CDMA plugins and dictionaries
  - ➔They are under the responsibility of each institute

Pratical details

- Technical contacts for CDMA
  - For SOLEIL :
    - Stéphane POIRIER poirier@synchrotron-soleil.fr
  - ANSTO
    - Tony LAM : tla@ansto.gov.au

- A Google group mailing list has been setup and is now the official place to share technical information

https://groups.google.com/group/common-data-model

Conclusion

As a temporary conclusion

- CDMA is a valuable response to the “Data Sharing” problem addressed in many European projects
- It is a solution that allows to deal with legacy files
- Implementation exists today in java
  - ➔even if project will benefit from having new software engineers looking at it
- Newcomers are now jumping on the boat
  - ➔We foresee to organize our first virtual CDMA meeting in Q1-2012 with all technical contributors

What would help us

- CDMA would benefit from being officially endorsed by 1 European project like Pandata
  - ➔*For example to convince DataAnalysis software developers (or companies) to adapt their software to CDMA*
- Adding some “man-months” of at least 1 experienced C++ developer
  - ➔*would speed up development time and give access to data to commonly used scientific analysis environments (like MATLAB or python)*