EDNA

A tool for online data analysis

and

A collaborative effort

Olof Svensson
EDNA Project Manager
ESRF Scientific Software Group

Overview

- Online data analysis
  - Example: MX automation
- The EDNA prototype
  - Data collection taking into account radiation damage
  - Results
- The EDNA Project / Framework
  - Project management
  - Data model
  - Modularity / plugins
  - The testing framework
- Future developments
  - The first EDNA release (Kernel + MXV1): May 2008
  - MXV2 / Tomography

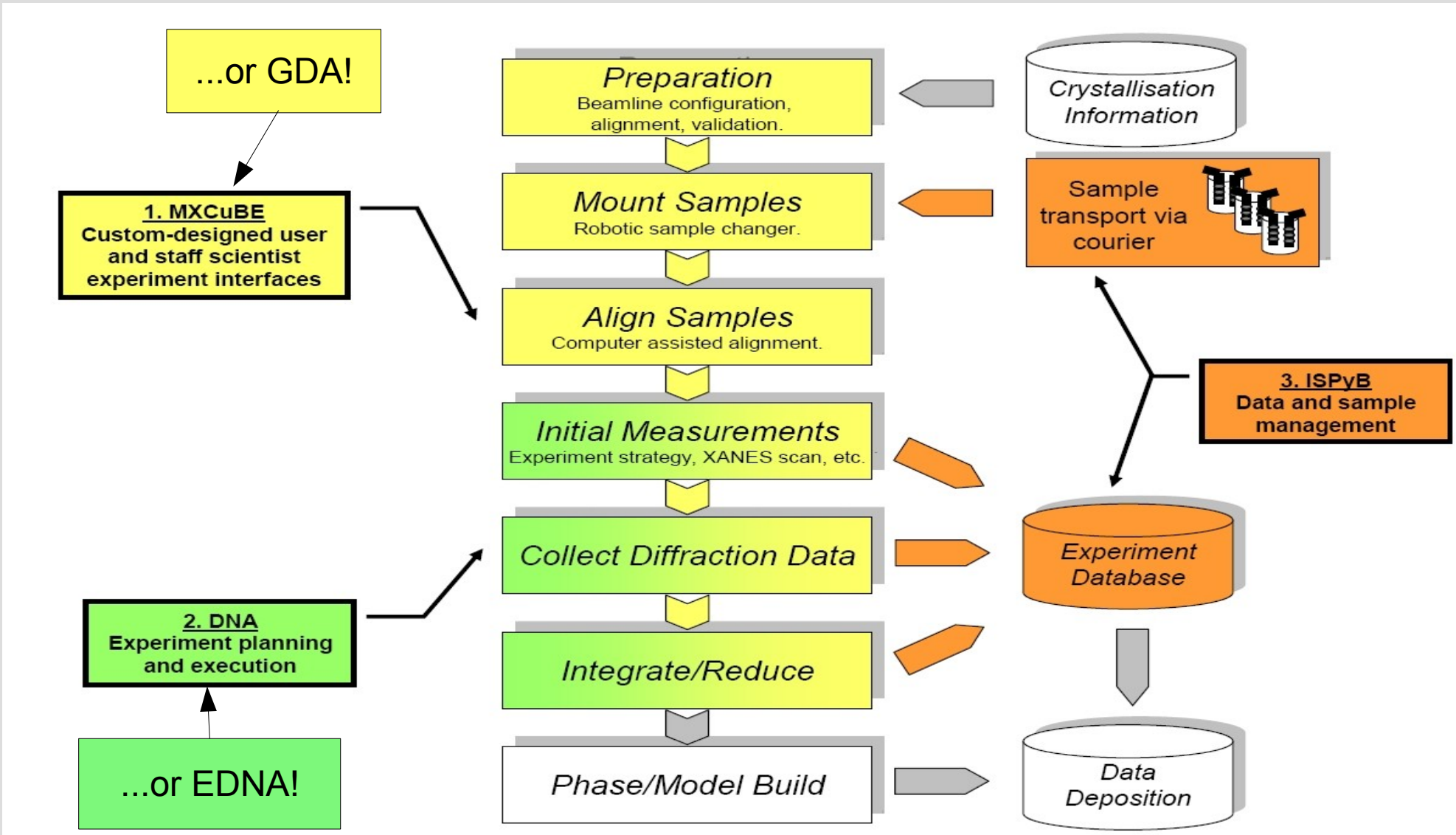
Online Data Analysis (1)

- Automation
  - Example MX : Data Collection Pipeline (DCP)
- Grid / Batch processing
  - Fast data processing $\Rightarrow$ parallel / distributed execution

Online Data Analysis (2)

- Data Management
  - Lims
- Modular
  - It must be easy to change the work flow / scientific programs
  - Data Model
- Robust – code reliability
  - If ODA fails once, the user won't try again...
  - Efficient error tracking system
- Collaboration
  - Avoid re-inventions of the wheel
  - Project management
  - Documentation, Use Cases

MX Automation - the Data Collection Pipeline: Full automation from sample loading to reduced (integrated and scaled) data

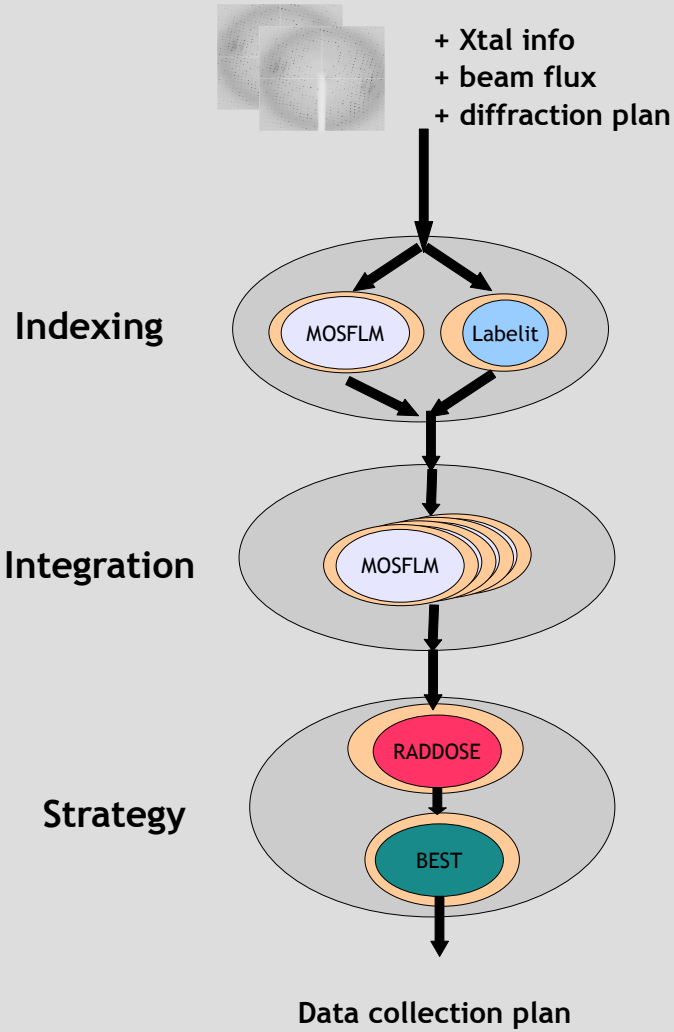


What is EDNA?

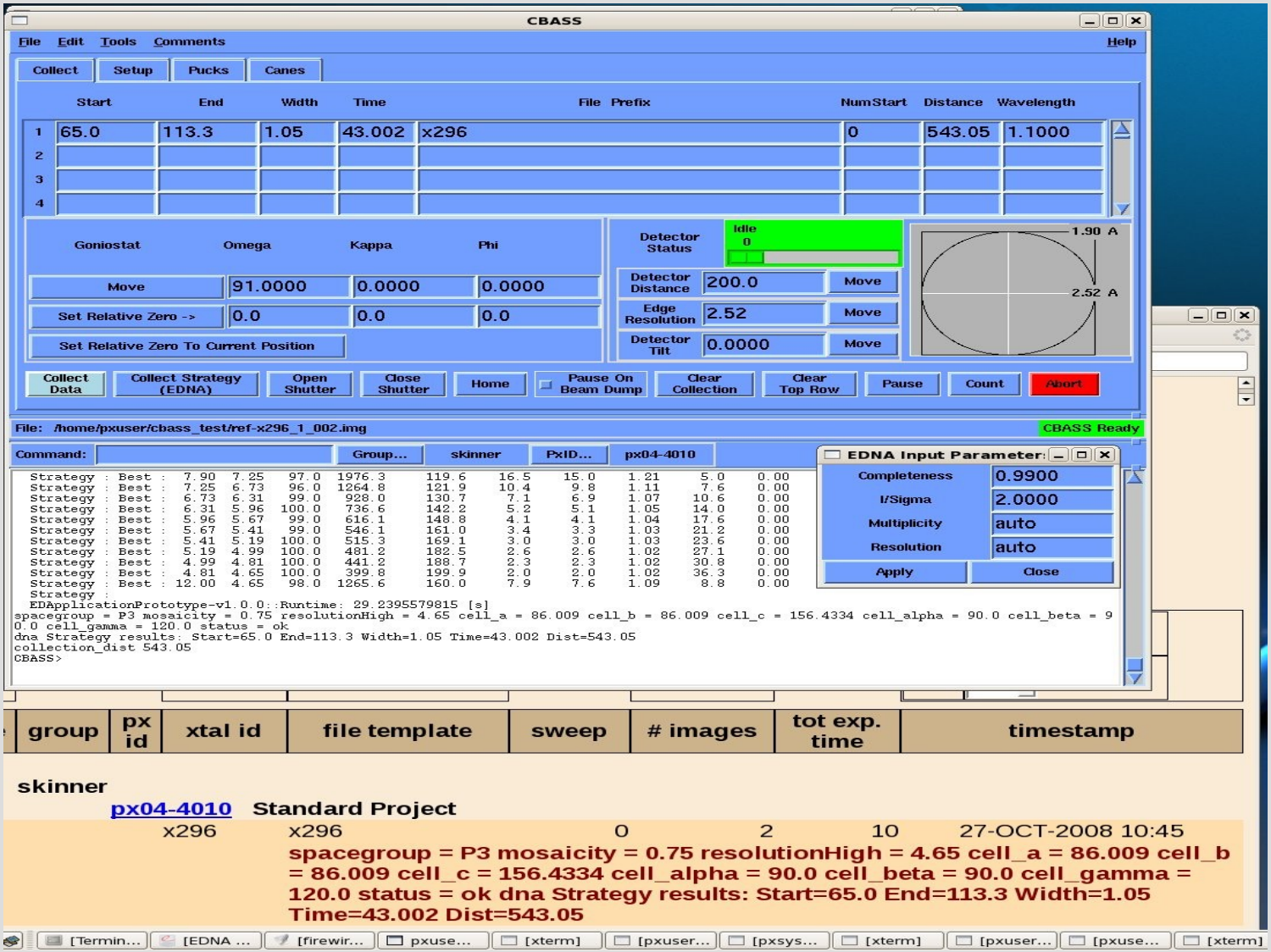
- EDNA is an international collaborative project between several institutes and synchrotron facilities.
- Developed on the foundation of the project automateD collectioN of datA (« DNA », www.dna.ac.uk)
- Designed to be a framework for Online Data Analysis of X-ray experiments

The EDNA Prototype

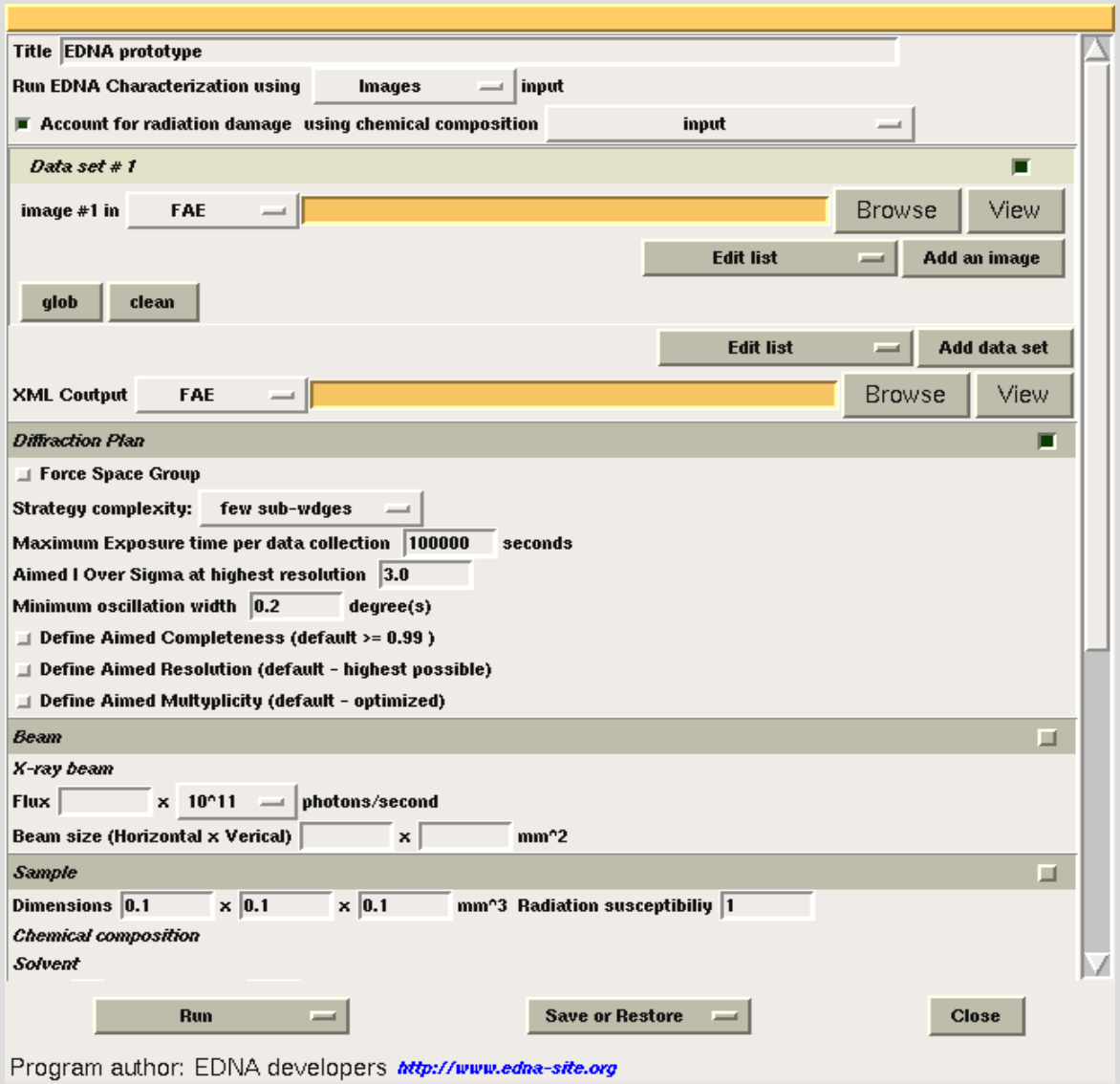
- MX sample characterisation taking into account radiation damage
  - Indexing using MOSFLM or Labelit
  - Parallel integration of reference images
  - If flux +beamsize + chemical composition: RADDOSE for estimating radiation damage
  - BEST strategy calculation
    - taking into account radiation damage
    - multi-subwedge data collection strategies



NSLS CBASS GUI



EDNA Prototype ccp4i GUI



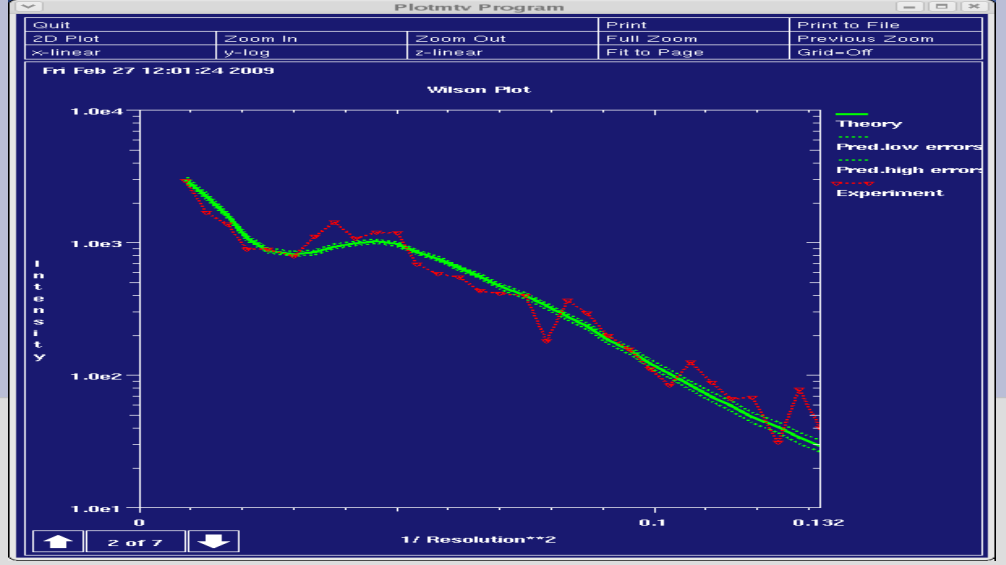
ID14-2, ESRF

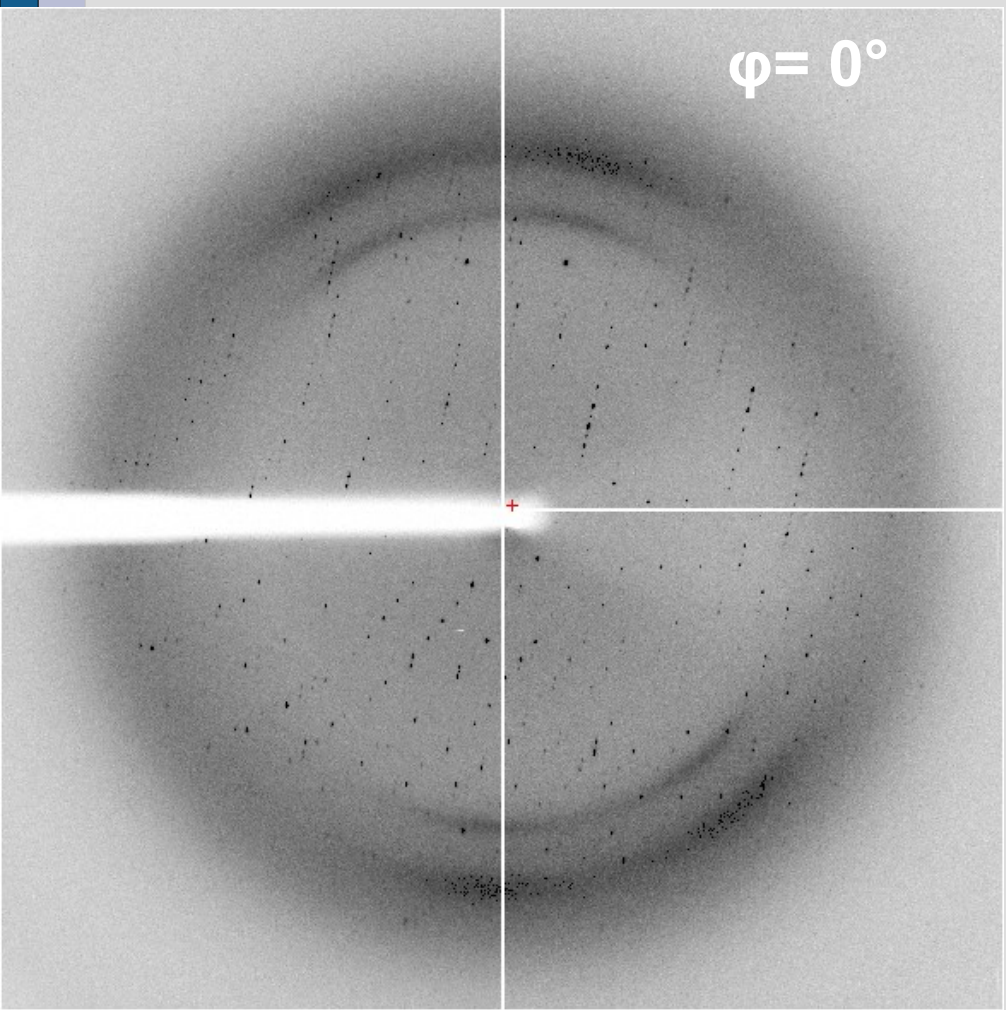
Initial images for Survival E protein crystal

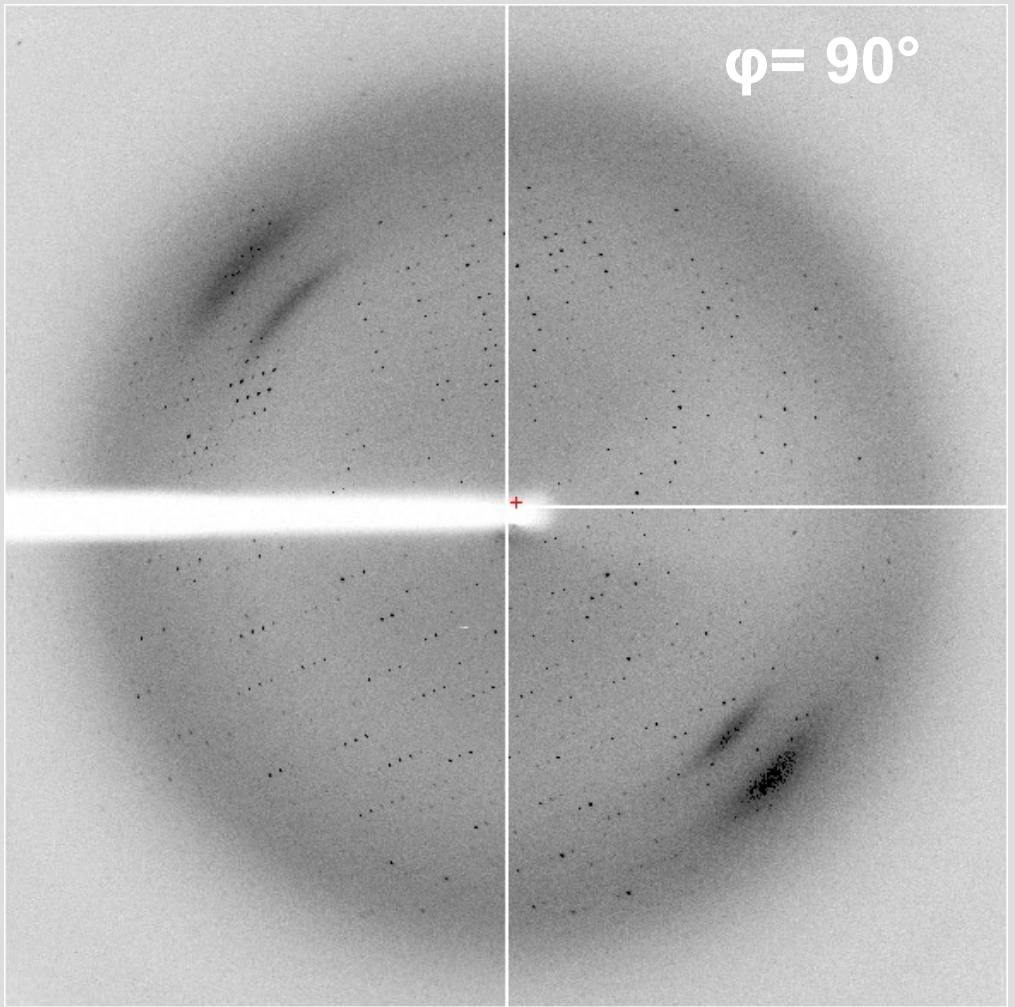
Exposure = 5 s,

Rot. width = 1°,

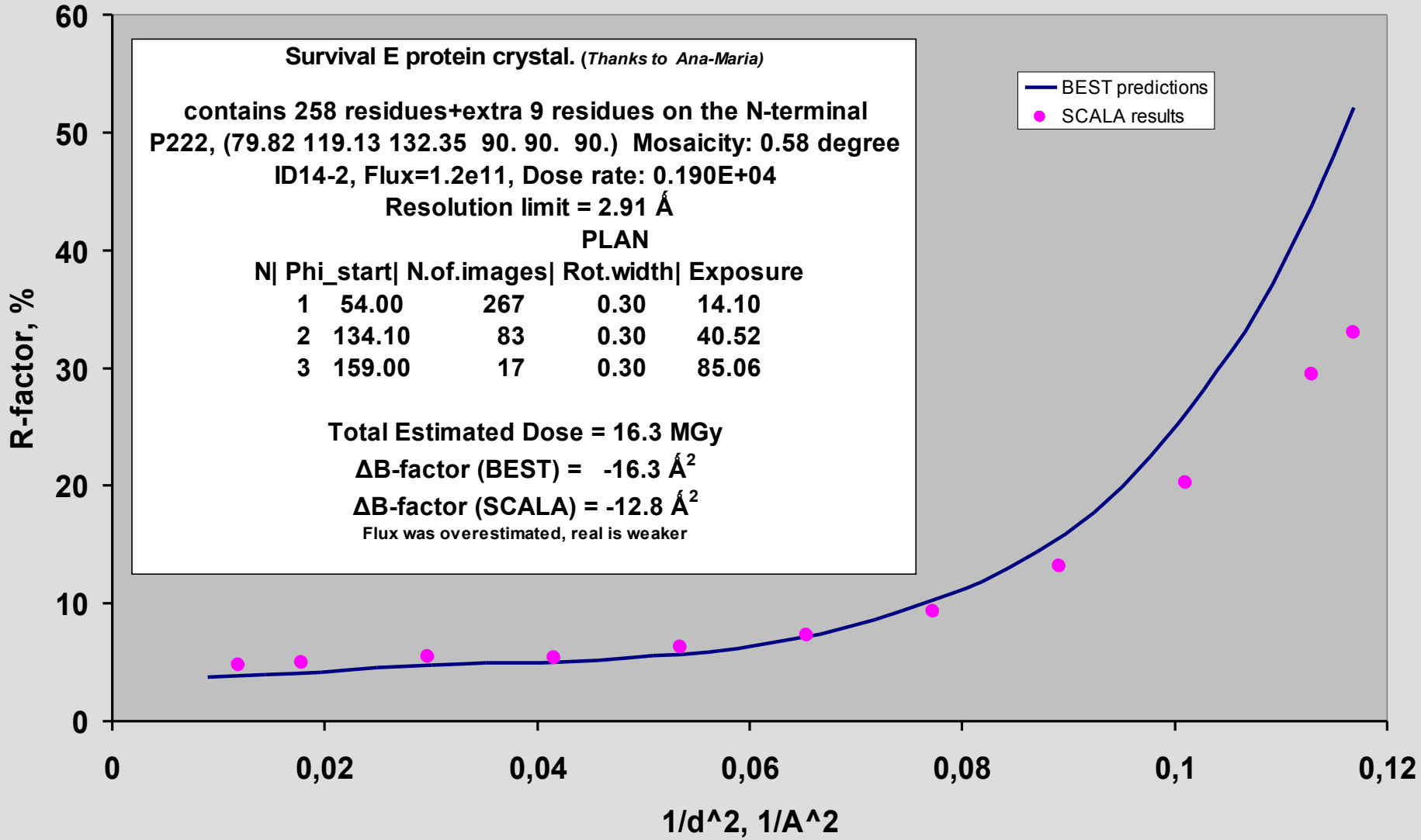
Resolution=2.8 Å



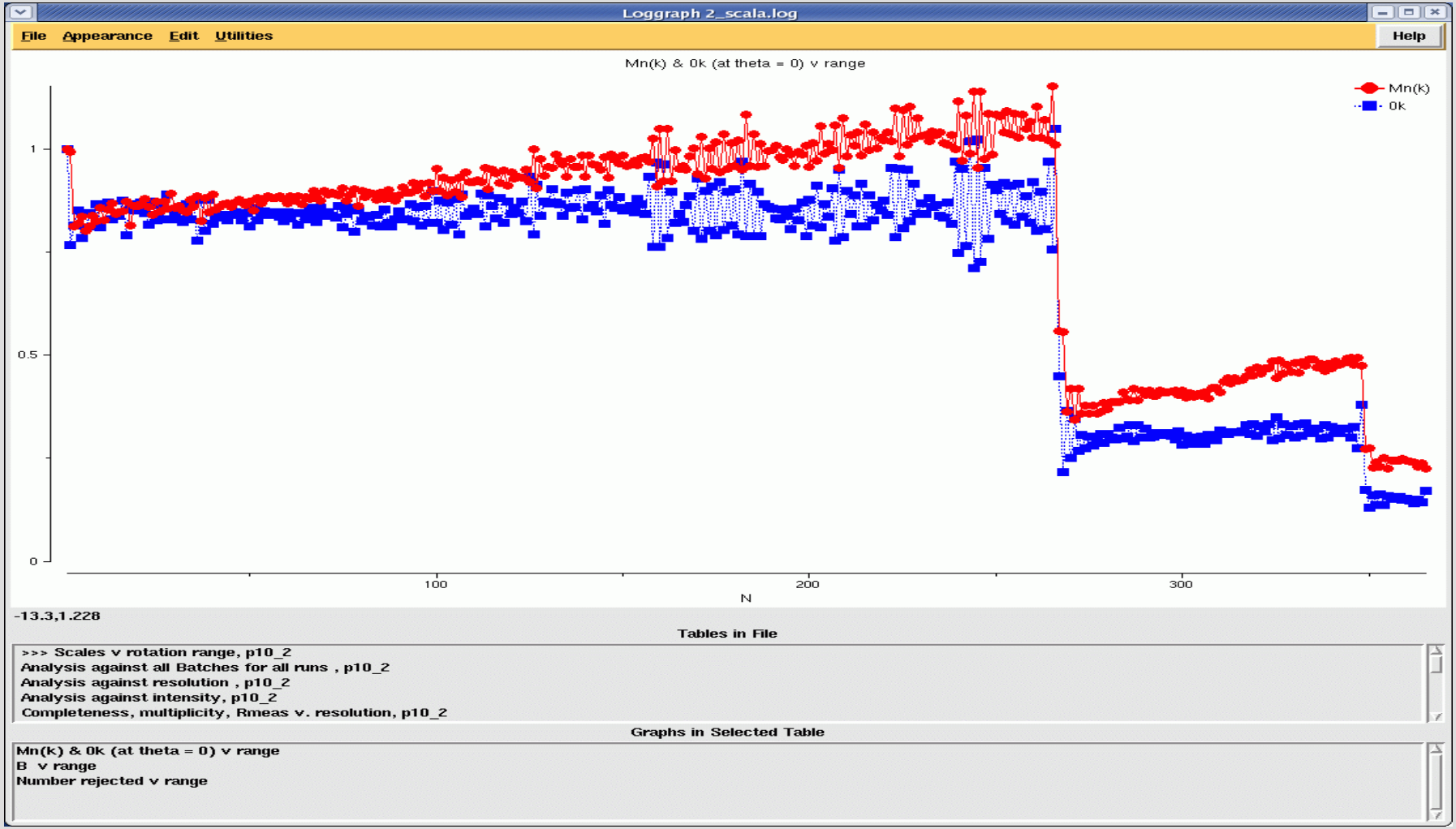




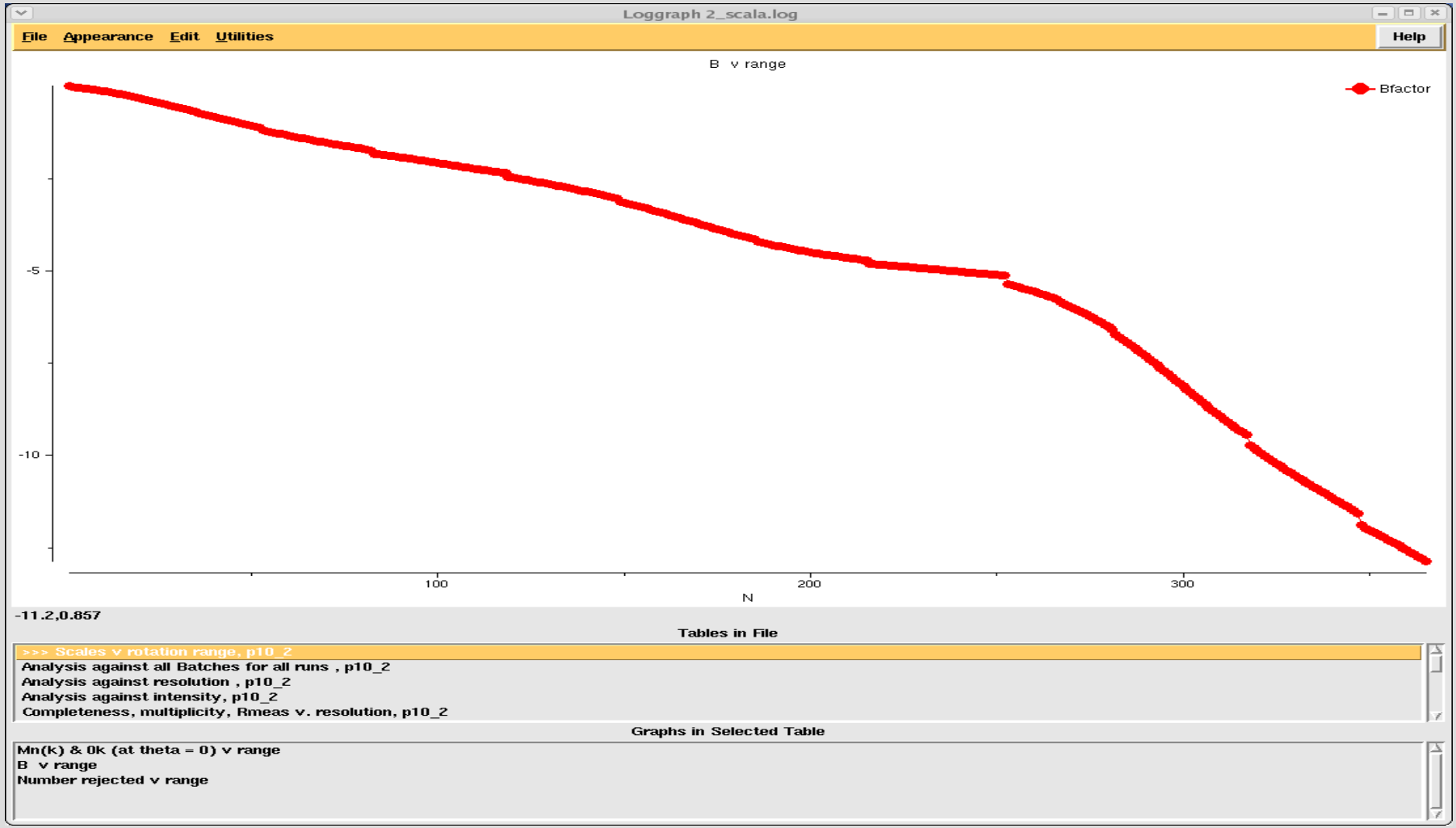
R-factor VS. Resolution



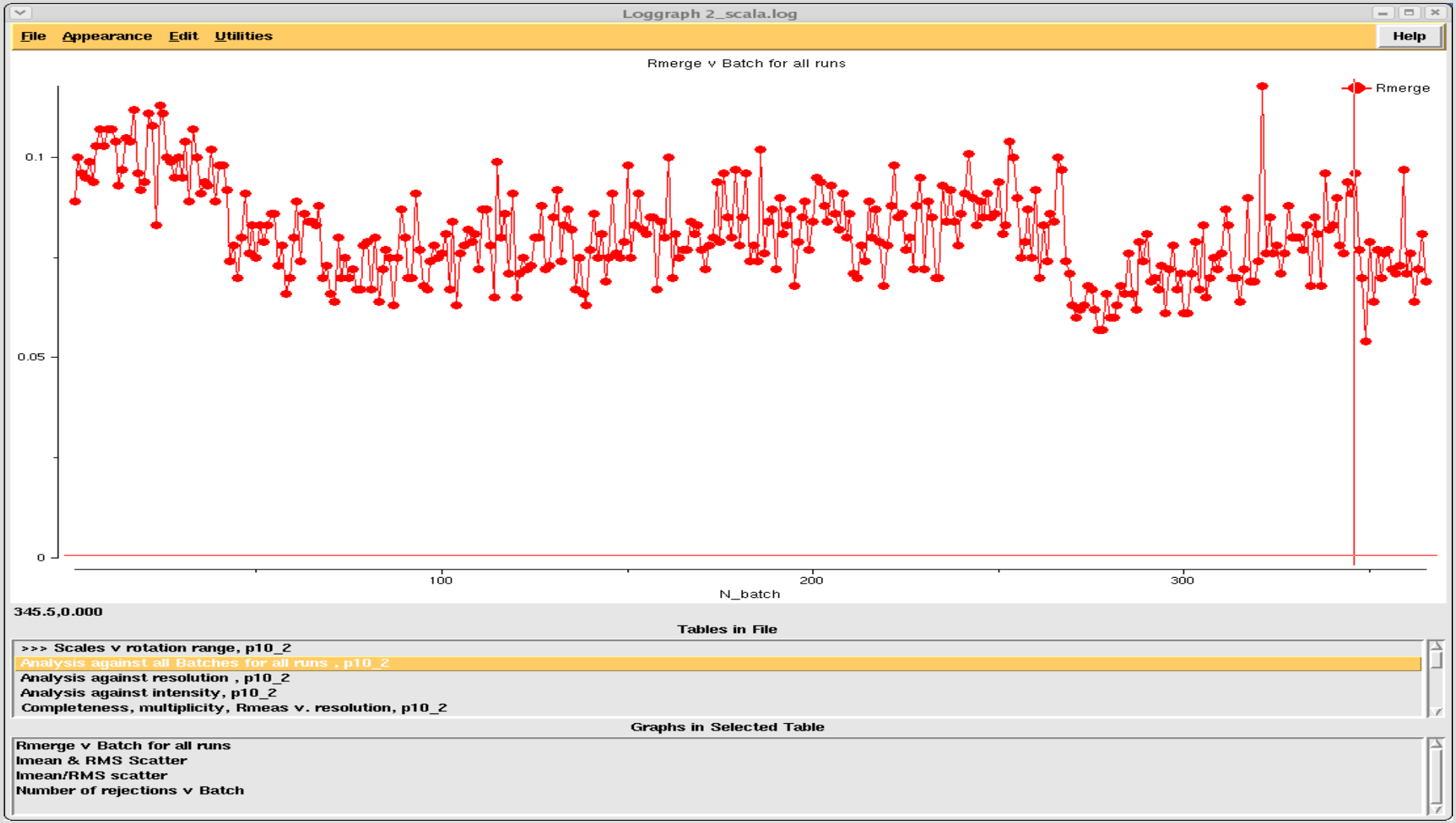
SCALA Output



SCALA Output

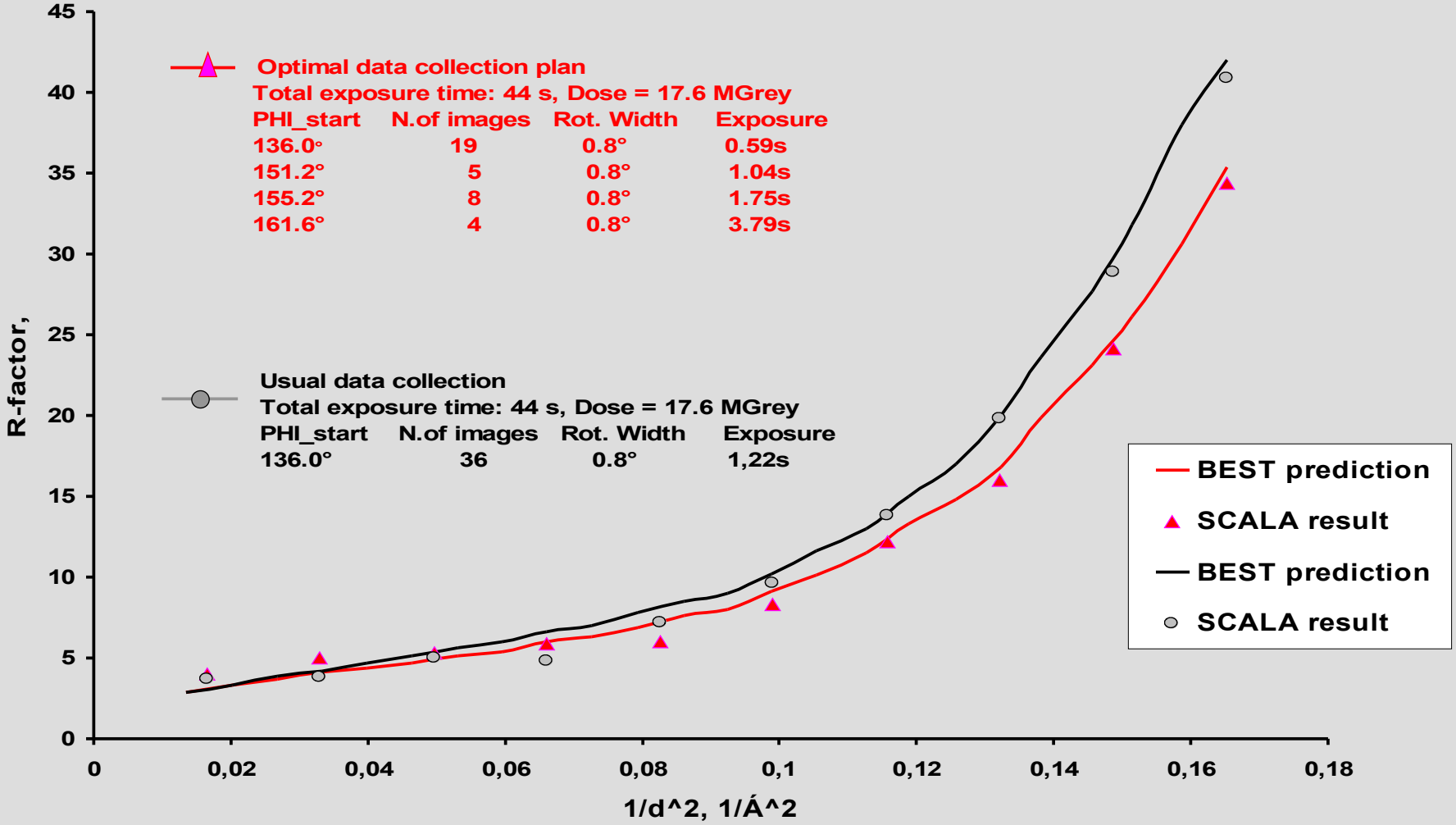


SCALA Output



ID23-1, ESRF

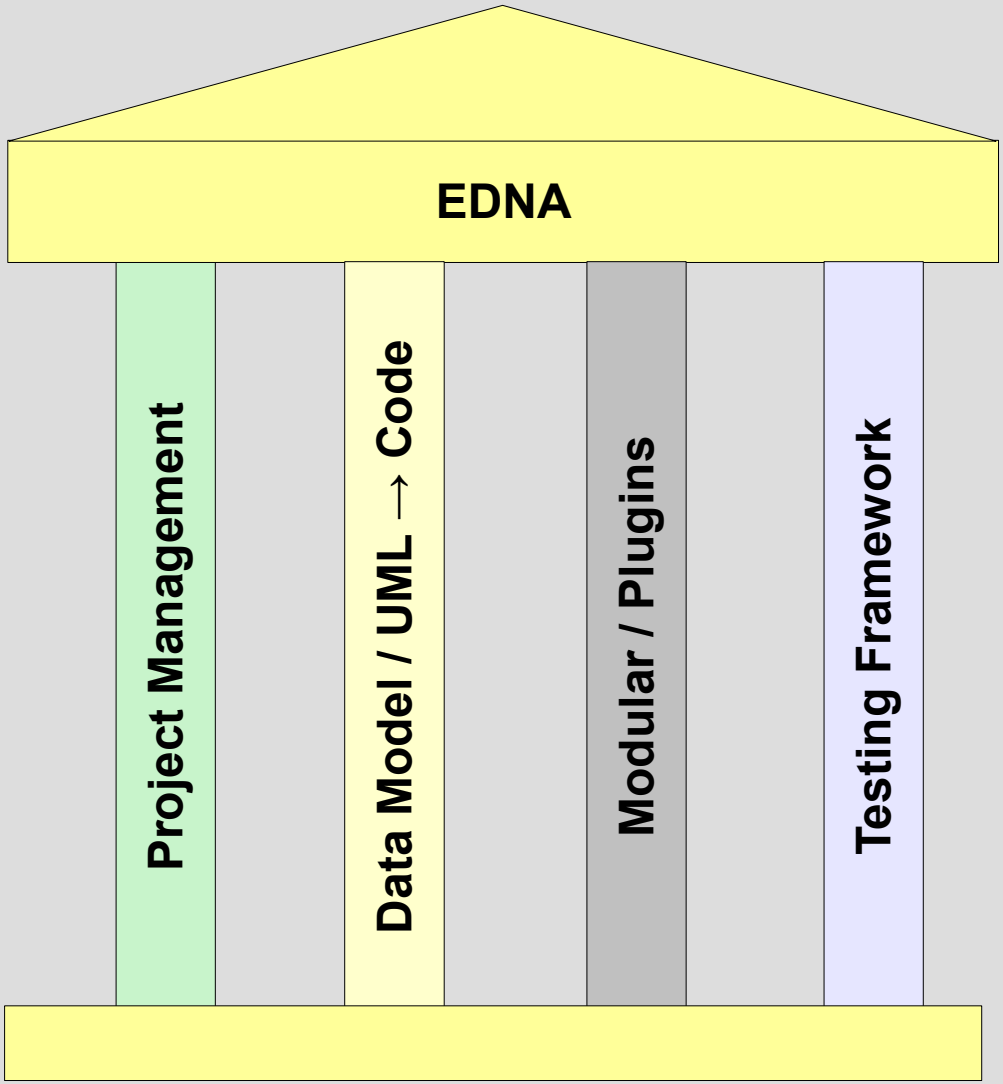
Complex of viral (from tomato bushy stunt virus, TBSV) supressor of RNA silencing p19 with mismatched siRNA



EDNA is not DNA...

- Same goals (initially)...
  - Automatic MX sample characterisation
  - Online data processing during MX data collection
  - Ranking
- ...however very different implementations
  - EDNA designed to not be specific to MX
  - No shared code base between DNA and EDNA
  - Different project management
  - Different collaborators
- If time allows:
  - Short presentation of DNA
  - History behind DNA → EDNA

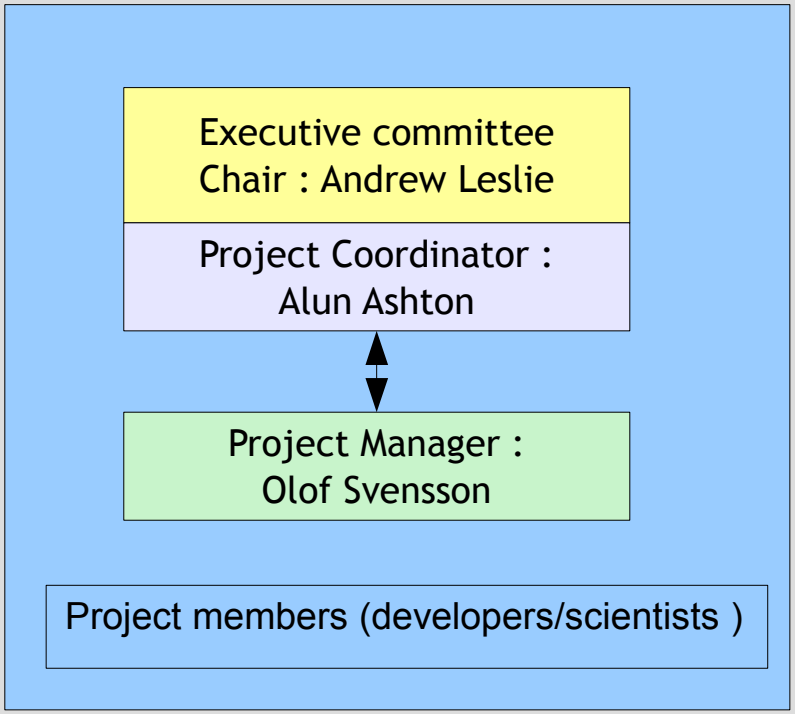
The EDNA Project / Framework



EDNA Project Management (1)

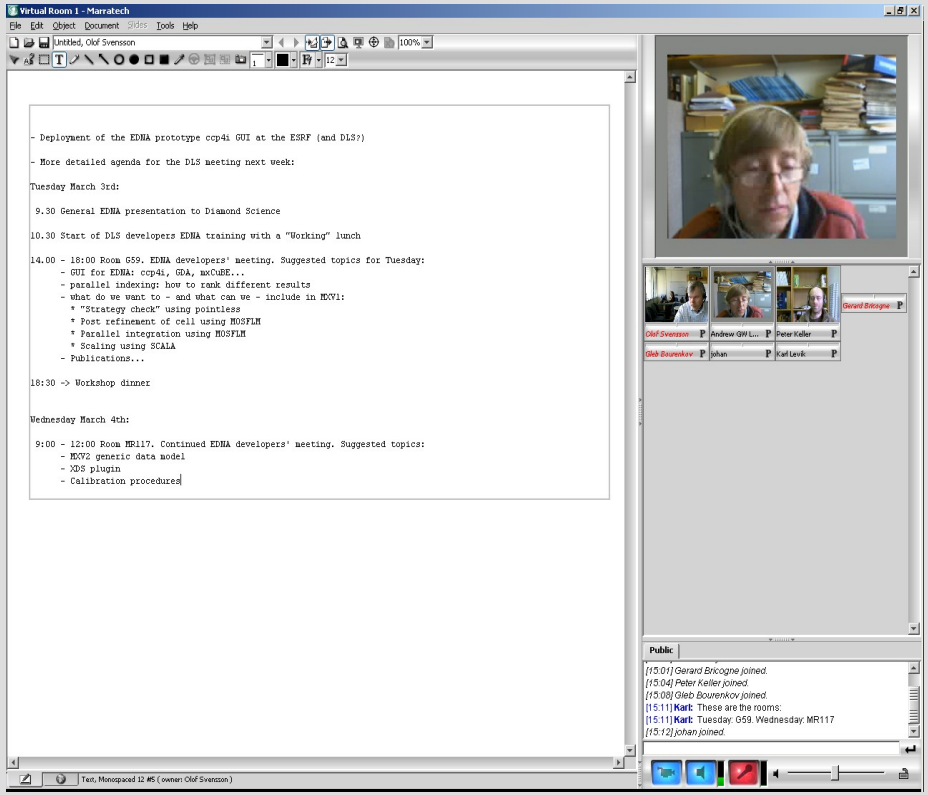
- Executive Committee:
  - Alun Ashton, DLS, UK
  - Gérard Bricogne, Global Phasing, UK
  - Andrew Leslie, MRC LMB, Cambridge, UK
  - Andrew McCarthy, EMBL-Grenoble, France
  - Sean McSweeney, ESRF, Grenoble, France
  - Thomas Schneider, EMBL-Hamburg, Germany
  - Andrew Thompson, Synchrotron Soleil, France

- Other members from:
  - BESSY, Berlin, Germany
  - MAX LAB, Lund, Sweden
  - NSLS,Brookhaven, U.S.
  - SLS, Villigen, Switzeland
  - University of Sydney, Australia
  - University of York, UK



EDNA Project Management (2)

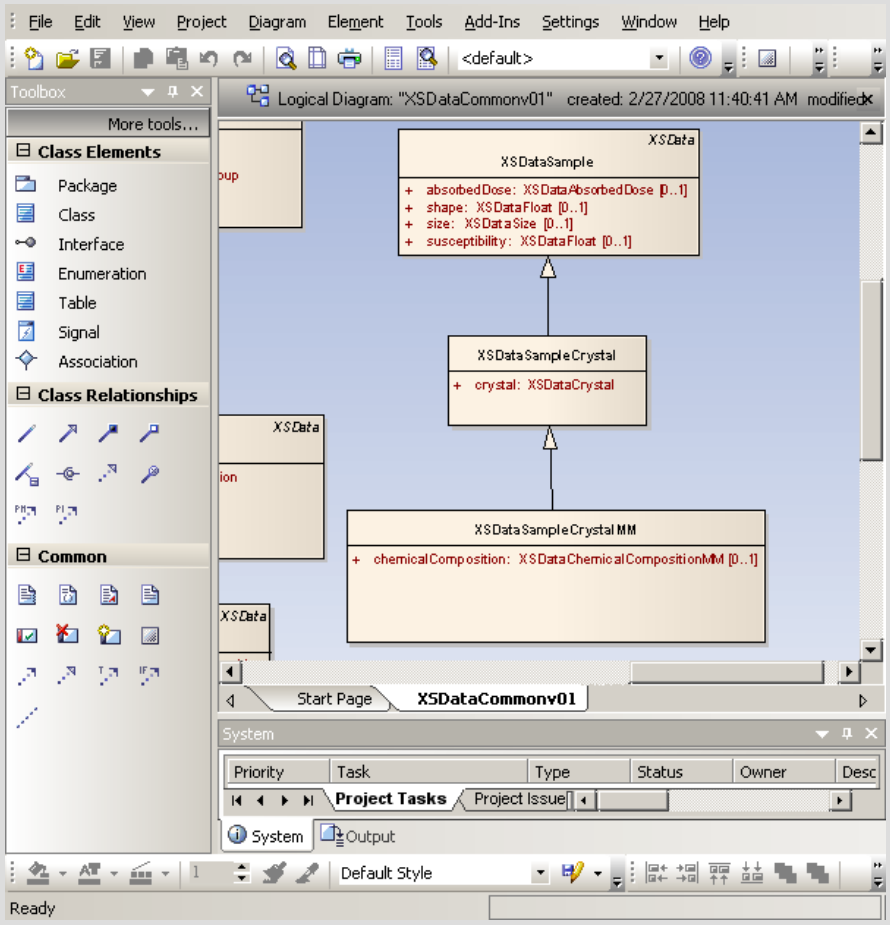
- Project agreement
- Coding conventions
- Code reviews
- Development tools
  - Eclipse
  - Enterprise architect
- Project portal
  - http://www.edna-site.org
  - Wiki documention
  - Bugzilla server
  - Subversion server
  - Discussion forum
- Executive committee
- Video conferences
- Developers' meetings & workshops



Marratech video-conferencing tool

The EDNA Data Model Framework

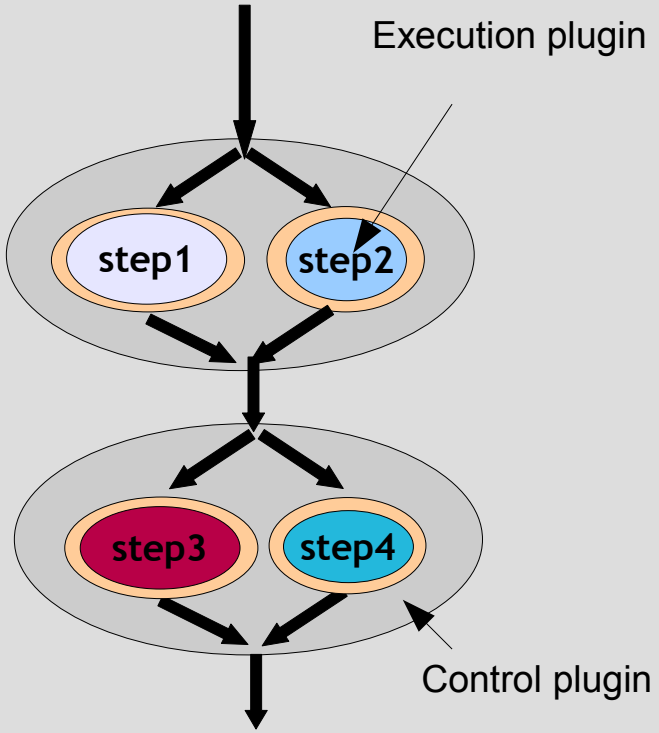
- From UML diagrams to generated code (data binding) :




```
<xs:element name="XSDataSample" type="XSDataSample"
<xs:complexType name="XSDataSample">
     <xs:complexContent>
          <xs:extension base="XSData">
               <xs:sequence>
                    <xs:element name="absorbedDose"
                    <xs:element name="shape" type="X
                    <xs:element name="size" type="XS
                    <xs:element name="susceptibilit
               </xs:sequence>
          </xs:extension>
     </xs:complexContent>
</xs:complexType>
```

EDNA Modularity : Plugins and their hierarchy

- Plugin base class :
  - Configuration, working directory, etc.
- Execution plugins :
  - Execution of external programs, e.g. (bash) scripts
- Controller plugins:
  - Control of execution plugins
  - Parallel execution
  - Synchronisation
- EDNA is based on AALib, however AALib is not a part of EDNA



EDNA Testing Framework

- The EDNA testing framework consist of three layers :
  - Kernel Unit tests
  - Plugin Unit tests
  - Plugin Execution tests

- Example of EDNA Plugin Execution tests result:

```
[SUCCESS] [ 1 ][ EDTestCasePluginExecuteReadImageHeaderv01.execute ][1.58020401001]
          [SUCCESS] [ 1 ][ EDTestCasePluginExecuteReadImageHeaderv01.testExecute ][1.41113901138]

[SUCCESS] [ 1 ][ EDTestCasePluginExecuteCCP4iv01.execute ][56.1572201252]
          [SUCCESS] [ 1 ][ EDTestCasePluginExecuteCCP4iv01.testExecute ][53.3898198605]

================================================================================


[UnitTest]: ############################################################
[UnitTest]: Summary Report:
[UnitTest]:                        Total TestCases: 17
[UnitTest]:             Total TestCases [SUCCESS]: 17
[UnitTest]:                Total TestCases [FAIL]: 0
[UnitTest]:                   [Total TestMethods]: 17
[UnitTest]:                               Runtime: 341.2 [s]
[UnitTest]:                                   Run: 00d:00h:05m:41s:239ms
```

EDNA Collaborators

Alexander Popov(d)
Alun Ashton(e)
Andrew Leslie(h)
Andrew McCarthy(b)
Andrew Thompson(k)
Clemens Schulze(j)
Clemens Vonrhein(f)
Darren Spruce(d)
Elspeth Gordon(d)
Ezequiel Panepucci(j)
Gérard Bricogne(f)
Gerrit Langer(b)
Gleb Bourenkov(b)
Gordon Leonard(d)
Harry Powell(h)
Johan Turkenburg(m)
Johan Unge(g)
John Skinner(i)
Karl Levik(e)
Katherine McAuley(e)
Lucile Roussier(k)
Marie-Farnçoise Incardona(d)
Mark Basham(e)
Meitian Wang(j)
Michael Hellmig(a)
Olga Roudenko(k)
Peter Keller(f)
Peter Turner(l)
Pierre Legrand(k)
Robert Sweet(i)
Romeu Pieritz(d)
Sandor Brockhauser(b)
Sean McSweeney(d)
Takashi Tomizaki(j)
Thomas Schneider(b)
Uwe Mueller(a)

(a) BESSY, Berlin, Germany
(b) EMBL, Grenoble, France
(c) EMBL, Hamburg, Germany
(d) ESRF, Grenoble, France
(e) Diamond Light Source, UK
(f) Global Phasing, Cambridge, UK
(g) MAX LAB, Lund, Sweden
(h) MRC LMB, Cambridge, UK
(i) NSLS, Brookhaven, U.S.
(j) SLS, Villigen, Switzeland
(k) Synchrotron Soleil, France
(l) University of Sydney, Australia
(m) University of York, UK

EDNA developers
Executive committee

Future developments

- Ranking in ISPyB
- Release of MXV1:
  - Based on the current prototype
  - Separate kernel
  - If time allows: implementation of data processing
- MXV2:
  - Improved generic data model
  - Implementation of XDS plugins
- Tomo:
  - EDNA for tomography

Why screen and rank?

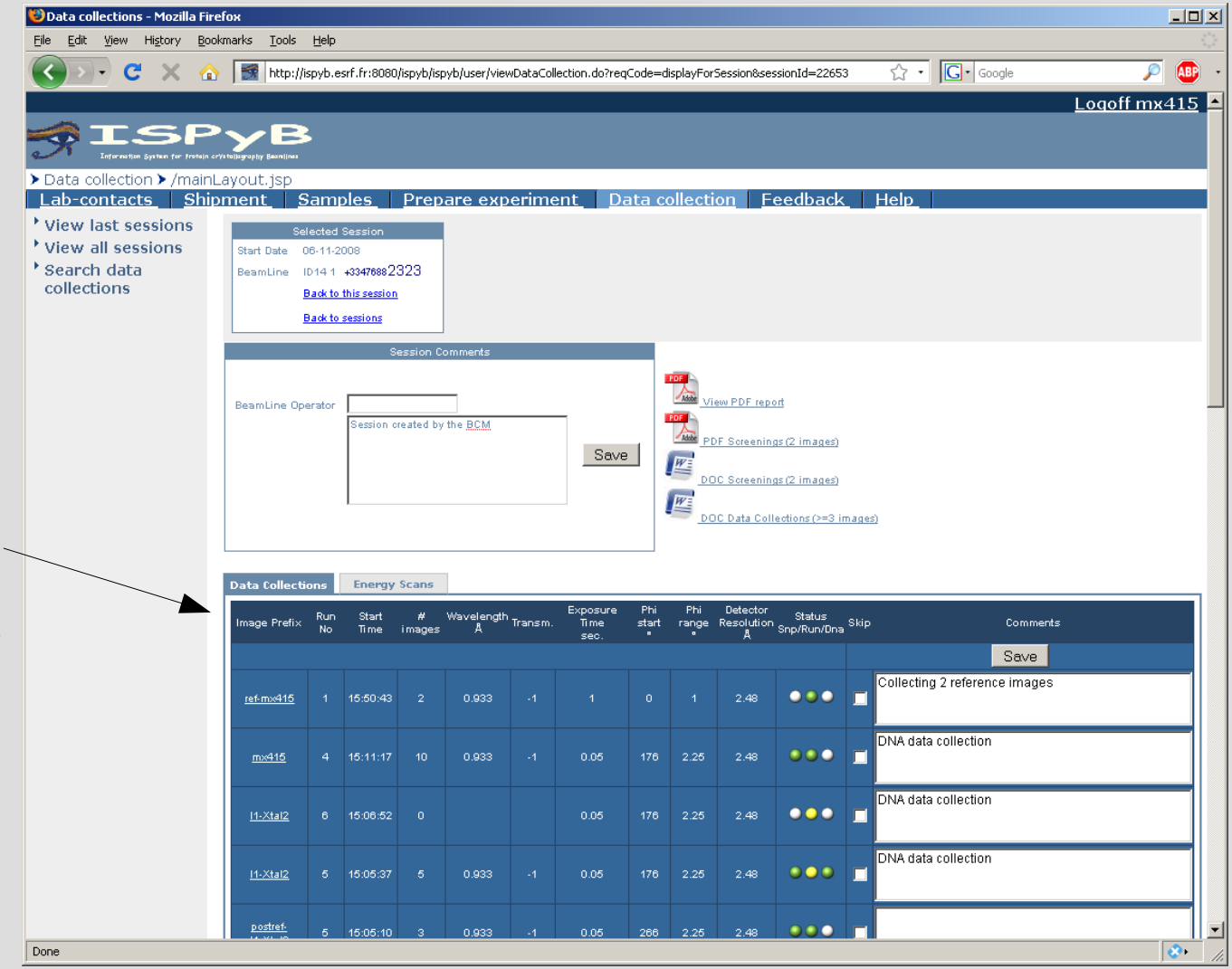
- On average at the ESRF : 24 data collections per PDB deposition
- Hence it's important to collect data on the best crystals
- Screening : collect reference images from a large set of samples of the same structure and calculate a score for each sample
- Ranking : choose the sample with the highest score
- Without automation : screening is time-consuming and tedious
- Thanks to automation, screening and ranking are fast and easy

Screening and Ranking with EDNA

- Screening of an individual sample corresponds to a single DNA / EDNA characterisation
- Different philosophy of EDNA compared with DNA :
  - EDNA has no control of data collection, hence no GUI for screening
  - Screening will be handled by mxCuBE and other beamline control GUIs
- No implementation of ranking yet with EDNA
- Current idea : Ranking via the ISPyB interface
  - Advantages :
    - Ranking can be made independently of data collection
    - Same interface for DNA / EDNA
  - Disadvantage :
    - No connection (yet) of ranking results to the data collection software (mxCuBE)

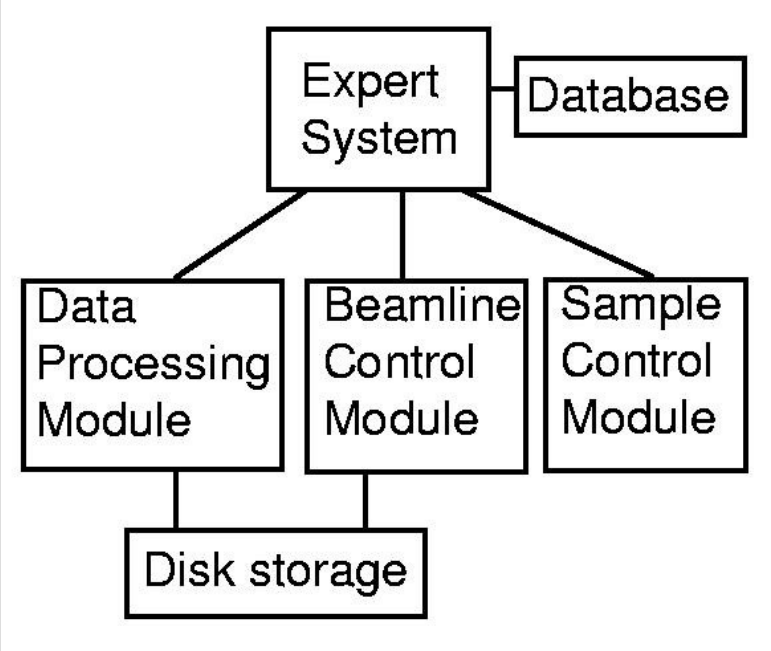
Ranking via ISPyB

- Data collection page in the ISPyB GUI
- If implemented :
  - Several different screenings should be selectable and form the input for ranking



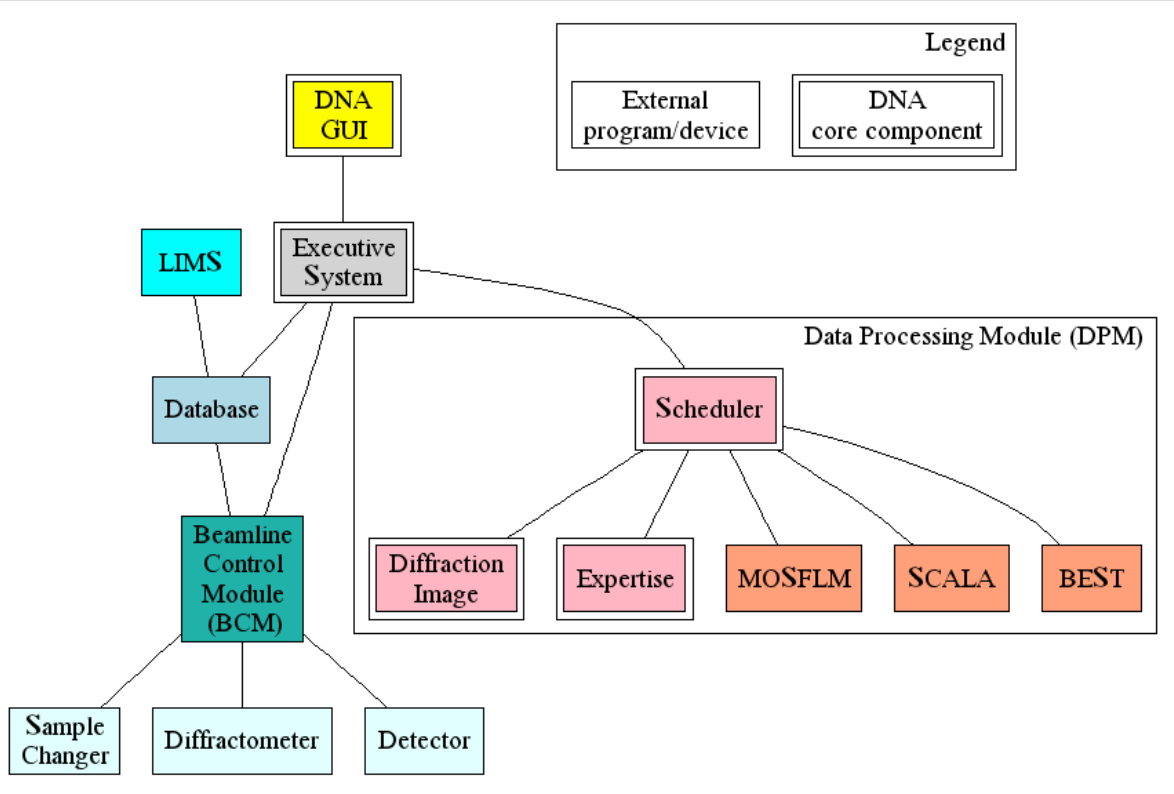
The DNA collaboration - the beginning

- Kick-off meeting in 2001
- Initial collaborators :
  - ESRF
  - Daresbury SRS
  - MRC LMB Cambridge
- Initially no external funding
- Meaning of ”DNA” :
  - automate**D** collectio**N** of dat**A**



Evolution of the DNA collaboration

- Main development period 2001 – 2005
- More collaborators and more developers entered the project, mainly thanks to external fundings : BioXHIT and e-HTPX.
- Installed and used at :
  - ESRF
  - Diamond
  - Recently used at NSLS (Brookhaven), now using EDNA
- Part of the 2008 BESSY Innovation Award



DNA collaborators

DNA Collaborators (in 2007)

Home institute	Name and link to email	DNA related work funded by	DNA tasks
Brookhaven Nat'l Lab.	Alex Soares	BNL	
	Bob Sweet		DNA co-ordinator at BNL
	John Skinner		
Diamond Light Source	Alun Ashton	DLS	Acting DNA project co-ordinator
	Colin Nave		
	Elizabeth Duke		DNA co-ordinator at DLS
	Karl Levik		DNA developer
	Katherine McAuley		Testing and implementation of DNA at Diamond Light Source
EMBL Grenoble	Raimond Ravelli	EMBL Grenoble	DNA co-ordinator at the EMBL Grenoble
	Sandor Brockhauser	BioXHIT	Working on introducing kappa geometry strategy into DNA
EMBL Hamburg	Alexander Popov	BioXHIT	DNA co-ordinator at EMBL Hamburg
	Gleb Bourenkov	DESY	Integration of the BEST strategy software
	Venkataraman Parthasarathy	SPINE	
ESRF	Sean McSweeney	ESRF	DNA co-ordinator at the ESRF
	Darren Spruce		Responsible for the ESRF BCM (ProDC) and the DNA - LIMS connection
	Olof Svensson		Responsible for the DNA Executive System and DNA 2.0 Project Manager
	Marie Francoise Incardona	BioXHIT	Working on DNA 2.0
	Romeu Pieritz		Responsible for developing the ranking module and working on DNA 2.0
Global Phasing	Gérard Bricogne	Global Phasing	DNA co-ordinator at Global Phasing
	Peter Keller	BioXHIT	Working on DNA 2.0
MRC LMB Cambridge	Andrew Leslie	MRC	DNA co-ordinator at Cambridge
	Harry Powell	CCP4	Responsible for the DNA DPM based on MOSFLM
MRC France - BM14 at the ESRF	Ludovic L auner	e-htpx	Working together with Darren on the DNA - beamline database connection
STFC Daresbury	Graeme Winter	e-htpx	Responsible for the DNA Scheduler
SLS - PSI	Takashi Tomizaki	SLS - PSI	DNA co-ordinator at SLS
Synchrotron Soleil	Andrew Thompson	Synchronton Soleil	DNA co-ordinator at Soleil
	Lucile Roussier		Working on DNA - database connection for Soleil
	Eric Girard		Working on offline tests of DNA
	Pierre Legrand		Working on integrating XDS as a DPM in the DNA system

What we got right in DNA

- Collaboration :
  - Scientists involved in design and testing
  - Executive committee for setting milestones / deliverables and for resolving conflicts between developers
- Two major Use Cases implemented :
  - Characterisation + data collection with online integration and quick scaling
  - Automatic screening and ranking
- Used regularly at DLS, ESRF and recently at the NLSL (now replaced by EDNA)

What went wrong in DNA

- The choice of name...
- Collaboration :
  - No project agreement
  - Minimal project management
- Not modular:
  - Too costly to change work flow
  - Poorly designed data model
  - Difficult for new developers to enter the collaboration
- MX hardwired!

A new Project!
Initial Design Thoughts

- A framework for Online Data Analysis
  - Use-case driven development
  - Modular – based on plugins
  - Configuration facility
  - Testing framework for assuring robustness
  - Data model tools and data classes code generation from UML
- Project management:
  - Executive committee, Project coordinator, Project manager
  - Code style / code reviews
- Application to other scientific online data analysis tasks
  - No MX specific code in the kernel

DNA 2.0 → EDNA

- The new project was accepted by the DNA executive committee Autumn 2005
- Project Manager
- DNA 2.0 officially launched in February 2007 in a DNA meeting held here at the DLS
- New name EDNA decided in the Project Agreement meeting at the ESRF in October 2007

EDNA Developments and Events

- February 2007 : Launch of DNA 2.0
- June 2007 : The spike meeting (ESRF)
- October 2007 : The Project Agreement meeting (ESRF)
- November 2007 : The first EDNA developers' workshop
- February 2008 : Developers' meeting at Soleil, France
- June 2008 : The prototype demonstration meeting (ESRF)
- August 2008 : Release of the EDNA prototype
- October 2008 : Plugin developers' workshop (ESRF)

Conclusions

- EDNA is a framework for online data analysis
  - Modular
  - Data modelling framework
  - Testing framework
- EDNA is a collaboration
  - Project management
  - Many facilities/institutes participating in the collaboration
- The EDNA prototype is ready to use
- The EDNA MXV1 application will be released in May
- MXV2 / Tomography applications are being planned

Acknowledgements

- The DNA team
- The EDNA team
- The ISPyB team
- Sasha Popov for providing me his slides
- DLS for inviting me and hosting this EDNA developers' meeting
- ...and thank you for your attention!