Grenoble, June 24th, 2013

From: E. Farhi, ILL
to: FP7/NMI3-II WP6 members and everybody interested by the topic

Data Analysis Standards (NMI3-II WP6)

3rd Meeting Minutes (June 20th 2013, Berlin)

Present:

- Emmanuel FARHI, ILL, France [farhi@ill.fr] +33 47620 7135
- Ricardo Leal, ILL [leal@ill.fr]
- Peter Willendrup, DTU, Copenhagen, Denmark [pkwi@fysik.dtu.dk]
- Jon Taylor, ISIS, UK [jon.taylor@stfc.ac.uk]
- *Joachim Wuttke, JCNS/FRM-II, Germany [j.wuttke@fz-juelich.de] (excused)*
- Thomas Gutberlet HZB, Germany [thomas.gutberlet@helmholtz-berlin.de]
- *Mark Johnson, ILL [johnson@ill.fr] (excused)*
- Martin Mueller, HZG, Germany [Martin.Mueller@hzg.de]
- Klaus Pranzas, HZG Germany [klaus.pranzas@hzg.de]
- Flavio Carsughi, MLZ/FRM-II [f.carsughi@fz-juelich.de]

Introduction
The meeting started with a reminder about the work package goals and resources, especially for the new attendees.

Task 1 : Review existing data analysis software and practices of software developers
The Task 1 report, which was published in March 2013 after approval and input from the work-package members, has been presented. It consists in an analysis of 24 existing and easily available software packages. The programming methodologies, user experience, data formats, as well as current practices are detailed. A list of recommendations is then produced. The report is available on the work-package web page <http://nmi3.eu/about-nmi3/other-collaborations/data-analysis-standards.html>.

Task 2 : Review existing solutions for a common data analysis infrastructure (mid November 2012)

The definition of a common data analysis infrastructure was discussed. The paper from G. Wilson "Best Practices for Scientific Computing" could be a starting point. It proposes a limited number of general rules to follow, but these do not guaranty a successful project by themselves.

The variety of available software provides a nearly complete set of solutions for most data treatments. A few identified missing functionalities still exist in e.g. off-specular reflectometry, t-o-f inelastic scattering and diffraction with crystal samples. A way to make use of these existing tools, while minimizing development and maintenance, is to re-use existing software as libraries with information exchange via memory, file or pipe. A proper rewrite in a new implementation may still be necessary when old projects are not maintained any-more and their internal complexity for further maintenance is beyond the development cost from scratch.

An other important point is that the size of collected data is only determined by the incoming flux, the scattering power of the sample (cross-section), and the detector solid angle coverage. Taking for instance 7. 10^5 n/s/cm^2 incoming flux, with a 10% scatterer (which is usual), and 0.6π sr detector solid angle, we draw 10^4 events/s to record. For 10^8 detector pixels, including time scale, we get an events record rate of about 24 kb/s, that is 90 Mb/h. Any significantly larger storage is attributable to noise. This implies that large detectors do not imply large data sets, and clever software design should allow support for any current instrument geometry with existing computers. However, building histograms with fine sampling will often require more memory.

The reason for the failure for large projects was also discussed. The diversity of scientists needs brings complexity with contradictory requirements, for instance the requirement for both raw data sets and processed data sets. Thus, this complexity is unavoidable as it arises from the diversity of instruments to support. An initially clear software design can thus evolve towards an unsorted complex set of routines, especially when the number of contributors to the project gets large.

However, it can be reduced for the user by a thorough documentation and for the programmer by the acceptance of standards for the definition of data, file, and method/calls. A suggested way to reduce software internal complexity is to define an integration team in large projects, which only commit new changes when they satisfy integration tests, proper documentation, and standards. This team acts as a reviewer layer, and should probably be composed of both trained programmers and scientists.

It was suggested to evaluate the use of modern software by making use of an automatic anonymous polling service which gathers which software/algorithm is used, with agreement from the user. This allows a better match between the user demand and the allocation of resources.

Action: write Task 2 report (E. Farhi and R. Leal with input from WP members).

Task 3 : Develop prototype software in chosen solution for representative applications

The Task 3 has been initiated end of 2012. The Mantid project has been selected for evaluation, as it is currently the only actively developed inter-facility effort. A learning period of 6 months was necessary before effective work could be produced.

As Mantid intrinsically supports pulsed data sets, *i.e.* time-structured, from the joint ISIS/SNS teams, we have agreed in previous meetings to focus on continuous source facilities. However, Mantid requires, in its actual version, time information stored in data sets in order for most algorithms to properly work. In this context, we have proposed to first focus our WP6 efforts on time-of-flight instruments located at continuous beam facilities, such as the ILL, MLZ/FRM-II, HZB/BENSC, LLB, and PSI/SINQ.

At the date of this report, data sets from 5 continuous source based time-of-flight spectrometers can now be loaded into Mantid (IN4,5,6 at the ILL, Focus at PSI, MiBemol at LLB). The existing Mantid 'algorithms' for data reduction have then been applied and compared with equivalent 'macros' in LAMP. It was found that in some cases Mantid is faster than LAMP (loading data), and in some others slower (transformation to $S(q,\omega)$) probably because of a non-optimal use/design of the corresponding algorithm to the special case of the continuous source t-o-f instruments for isotropic materials (|q| mean). Also, Mantid requires significantly more memory to store data sets than LAMP, as the time axis is duplicated for all pixels composing the detectors, whereas some other packages (e.g. LAMP) prefer a unique definition shared by all detector pixels. All in all, we find that results from Mantid and LAMP compare well.
A similar effort is under way to support the TofTof spectrometer at FRM-II, and the D33 SANS-ToF at the ILL.

We encourage work-package members to install the latest Mantid release at their home institute, and in particular on the newly supported instruments when applicable. The storage in NeXus format is highly recommended.

Action: the case of multiplexed TAS and powder/SX diffractometers could be the next step in the Mantid evaluation.