Planning and organization of an e-learning training program on the analysis software in CMS

Kati Lassila-Perini¹, Sudhir Malik², Benedikt Hegner³, Andreas Hinzmann⁴ and Roger Wolf⁵

¹Helsinki Institute of Physics, ²Fermilab/University of Nebraska, ³CERN, ⁴RWTH Aachen University, ⁵University of Hamburg

kati.lassila-perini@cern.ch

Abstract. Analyzing physics data at LHC experiments is a complicated task involving multiple steps, sharing of expertise, cross checks and comparing different analyses. To maximize physics productivity, the CMS experiment at LHC has developed a collection of analysis tools called the PAT (Physics Analysis Toolkit). A comprehensive training program was designed and setup on using PAT software as an integral key part of the analysis of data from the CMS experiment. This note summarizes the requirements and the considerations that led to establish the PAT training program, the approach taken, the main organizational issues and the experience on its implementation and maintenance. The training and the feedback has proved to be very successful both for participants and the PAT development team.

1. Introduction

The professional and educational background of software tools users in the high energy physics experiments such as CMS[1] is often not enough to efficiently perform data analysis. The tools are intrinsically complex and very domain and experiment specific and therefore the skills acquired in the university courses are not sufficient for a quick start. Traditionally, skills and knowledge specific to an experiment have been passed from seniors to juniors within a research group of an institute. However, during the intense software development phase at the start of the data-taking, not all members of the group have the required skills and therefore the traditional senior-junior mode of passage of knowledge may be hindered.

The sheer size and geographical spread of the large international research collaboration require new ways of spreading the experiment specific knowledge. Many advanced tools are already in use. The information and documentation is available on the internet. The meetings and presentations can be transmitted (live or recorded) to different geographical locations with well established tools. The training based on these elements, available to all members, allows for a democratic access to data and knowledge independent of the member's location.

To tackle the training challenge, many short tutorial sessions have been organized on computing and software tools, including the general CMS software framework[2] and the Physics Analysis Toolkit (PAT)[3]. They have been well attended and appreciated, but short sessions did not allow users to put in practice the skills they acquired in the context of their own analysis. It was also difficult to address the advanced features as the previous knowledge of the participants was either not known or known to be very heterogeneous. Finally, the real learning did not happen during the tutorial session.
but afterwards, individually, with the help of questions to discussion forums, questions to experts (or non-experts) or copying and pasting from examples that may be of uncertain quality. These considerations made obvious the need for a comprehensive training that results in a good quality software using PAT and to an adequate expertise on the use of the tools. The e-learning project described in this note addresses the above challenge.

Organizing a comprehensive training is not a negligible task to a software development team. Therefore, the profit of such training to the team has to be assured. The main motivations and goals to the PAT development team are i) getting feedback, ii) training experts, iii) improving documentation and iv) teaching efficiently. These goals have indeed been achieved with the training. Eight instances of the course have been organized and more than 200 CMS physicists have been trained.

2. Organization
A project plan was set up with a detailed definition of the course – goals, format, main contents, the duration and the main actors of the project team and their roles. The latter are

- project manager – ensures the timely execution of the project plan
- responsible teacher – assigns the lecturers and tutors and overviews the coherence of the course-ware
- content providers – provide and update the exercise material for the teaching unit assigned to them
- lecturers – provide and update slides for the teaching unit assigned to them, give the lecture
- tutors – are present during the lecturers and answer to the students' question
- local organizer – books the rooms and takes care of the video transmission and recordings

The formalized approach of the project plan was found to be an excellent support when discussing the project in the beginning and helped to clarify the responsibilities, especially in the case where some actors know each other very well and others not at all. In practise, one person can be a content provider, a lecturer and/or a tutor for one or several teaching units and more than one person can provide content to a single teaching unit. It is important, however, that the tasks of the project manager and the responsible teacher are not shared although a frequent consultation among all key actors is and should be a common practise.

A detailed work-flow is written for each instance of the course. It defines in detail the specific tasks taken over by each member of the project team. The written and agreed work-flow makes the organization of a new instance of the course fairly straight-forward.

3. Pedagogical approach
The goals and contents were defined together with the Analysis Tools group conveners and the PAT development team leader. For each unit, an expert (content provider) was assigned to provide material. These units cover all areas from basic knowledge to the details needed in a real analysis.

To assure efficient learning, all participants are required to have an a project in mind to which the skills acquired during the course can be applied. This ensures that the participants have an immediate use-case for all information. In this stage of their professional career, the participants are not interested in an academic exercise but they want, instead, the training to lead to software skills which are easily customized for their own research projects.

Each unit has a set of exercises and activities. The “first principles of instruction”[4] were found to be directly applicable for our training. Therefore, we have set the following goals:

- demonstration promotes learning
  - training material should include a demonstrative example
- applying new knowledge promotes learning
  - training material should include a guided exercise in which the participants will reproduce the example and modify it slightly
- activating prior knowledge promotes learning
• the content providers should know what items the precedent units have covered and include exercises requiring the use of these skills
• integrating new knowledge to the everyday world promotes learning
  ° training material should include an activity which connects the newly acquired skills to a real-life situation
• task-centred instruction promotes learning
  ° training material should be designed in function of the skills needed to complete the analysis software project of each participant.

In punctual training sessions, only the first and rarely the second of these principles can be applied. The current course-ware is designed keeping these principles in mind and it has indeed promoted efficient and engaging learning.

The training was chosen to be a mix of physical presence and remote learning. The remote participants in the European time-zones follow the units with lectures through a videoconferencing system and those in American and Asian time-zones can watch the recording of the session which is attached to the agenda after each unit.

In the high energy physics research environment, usage of collaborative internet tools for remote participation has a long tradition, and the innovation - somewhat in contradiction with most e-learning training projects - lies in adding more formal components to the informal training methods already in use. In practice, the main elements to improve the efficiency of learning are:
• limited number of participants
• limited and defined length of the training
• close follow-up of the progress on exercises given during the course

All these elements are normal factors for a standard university course, but in the research environment they introduce a certain innovation in the training and in the user support.

4. Learning environment
A learning web platform was built for the training. The platform contains
• announcements and general instructions
• link to the course schedule
• link to the course material
• activities and exercises and exercise submission area with an access limited to the participants
• discussion forum

A collaboration workspace area centrally provided by the CERN IT department was chosen for this purpose. The advantages of this choice are the predefined graphical layout and the possibility to customize the area in function of the specific needs for the course.

4.1. The course material
The course material (slides, exercises, recordings) is required to be accessible to the full research collaboration and therefore they reside in the usual area for presentations and documentation - CERN agenda server for the slides and Offline Software Guide for the exercises and complementary reading material. The agenda server also offers functionalities such as registration and mailing list to all participants. The procedures specific to each course instance are defined in the course platform.

4.2. The job submission and feedback forms
Forms for the exercise submission and for the final feedback are defined in the web platform. The participants can view and modify only their own forms while the training team can view them all. The main goal of the exercise submission forms is to allow the training team to follow the progress of the participants. The correctness of the response is only of secondary interest.

4.3. Administration area
The learning environment includes an administration area that is used during the preparation of the course. This helps the training team to get familiar with the environment and helps customizing the area according to the needs of the course. For the first instance of the course, which required a longer preparation time, the administration area included a planning form for each module that were filled by the content providers with outlines of the issues they plan to handle in their module. For the following courses – apart from some restructuring from time to time – such a detailed planning area is not needed.

5. Standard course format
The standard course format (figure 1) has been established as a one-week course preceded by a set of prerequisite exercises. The prerequisite exercises are given as a self-study material to be completed during two weeks prior to the course. The tutorial week consists of lectures, discussion and presentation of the homework exercises during the morning session. The afternoon is left for the participants to work on the exercises on their own. The participants submit replies to the exercises on a web form so that the training team can follow the progress and verify that exercises are actually being done. A discussion forum is available to the participants in case of problems and questions.

This format was achieved after several training courses taking into account
- the difficulties encountered by the participants
- the feedback from the participants after the course
- the feedback from the tutors and the lecturers
- the organizational constraints such as availability of the tutors and the lecturers.

![Figure 1. Standard course format.](image)

5.1. Prerequisite exercises
The use of PAT tools requires a solid knowledge of the underlying software and computing environment. During the first rounds of the course, it was observed that a vast majority of the problems and uncertainties did not raise from the use of PAT itself but from various sources connected to the general CMS software framework. Therefore a set of prerequisite exercises was added to bridge this gap and it was defined as obligatory for participants with no or little knowledge on CMS software.

The prerequisite exercises cover topics such as
- simple copying-and-pasting from the instruction page to the working environment
- setting up a CMS software release area
- finding data
- working with the source code repository
- using common CMS software framework utilities
- understanding the basic features of the build environment
- using tools to browse and search in the source code
- building a basic application in the CMS framework, making histograms
• understanding the basic features of a CMS software python configuration file for a simple analysis job
• accessing the grid environment

The exercises include tasks or questions on each key point. Links to further information in the common CMS software documentation suite are listed thus promoting the awareness of the existing knowledge base.

The exercises are made available two weeks before the start of the course, and the participants are informed that they should submit their replies on a web form. It usually requires a close follow-up and several reminders to get everyone started, but it pays the effort in terms of learning efficiency. The time during the course can then be used in learning the subject itself. The participants have found these exercises useful and a vast majority of them are able to complete them before the course starts.

5.2. Lectures
Each unit is introduced in a lecture of roughly one hour. The lecture explains the physics use case for the topic in question and gives some demonstrative examples. The lecture units are
• introduction to the Physics Analysis Toolkit and the motivation for its use
• finding information and solving problems using PAT and CMS software documentation
• producing data in PAT format and accessing them
• configuration of contents of a data file
• embedding additional user-defined information
• disambiguation of physics objects using PAT
• tools for Monte Carlo truth matching
• using and analyzing trigger information
• full-scale physics analysis example.

The lecture material resides in a common slide pool where a lecturer can pick up the most recent version of the material and make the necessary updates adapting for software version changes and for the new features in the software itself. The slides are designed in a way that they can be used in self-study.

The lecturers are encouraged to rotate the topics from time to time. Furthermore, the tutors are encouraged to become lecturers on the topics where they have gained or wish to gain enough experience. Such recycling promotes the expertise in the training team and makes the organization less vulnerable in case of unforeseen absences.

5.3. Tutorial exercises
Each unit is completed by a set of tutorial exercises. The exercises consist of a fairly simple task where an existing example is modified and of a more involved task where the example is further customized following the use-case given in the instructions. In the latter, the skills acquired in the previous units are usually needed.

The participants are requested to submit a reply to a set of questions indicated in the exercise page. The replies are submitted on a web form on the learning web platform. Some of the questions are simple and are meant to stop the reader and require some thinking and other questions are numerical values or object names which indicate that the exercises are done. The introduction of such web form has clearly increased the participants’ motivation to go through the exercises – such small details can make all the difference between “just reading” and “doing” and an equal difference in the learning efficiency. Similarly, as with the prerequisite exercises, a close follow-up is needed and the participants are reminded to do the exercises and submit the answers if needed.

The exercise material resides in the common CMS software documentation suite. Each page has a responsible person (often also a tutor or lecturer) who makes sure that the material is updated before each course. Another person (often a tutor) goes through the exercises and makes sure that the instructions are complete.
5.4. Tutoring
Tutoring is an integral part of the course and the participants can expect to get help while working on the exercises. The tutoring is mainly done through the discussion forum in the course web platform. During the first rounds of the course, hands-on sessions were organized, but they were abandoned due to inherent problems: different working pace of the participants makes efficient work difficult; the meeting rooms often suffer from an insufficient wireless connectivity; and it is hard to measure how far the participants have progressed in the exercises.

5.5. Feedback
After each course, a detailed feedback form is made available and all participants are requested to submit their feedback. The training team meets informally and summarizes the observations from each course. Furthermore, the training team follows to the extent to which the participants have been able to complete the exercises. This feedback is extremely valuable when evaluating the course and has lead to several well motivated changes and improvements.

6. Conclusion – key to success
The eight course instances in two years of the PAT e-learning training have been a success. More than 200 physicists have been trained. The documentation has been restructured and improved and it is constantly updated and available to the full collaboration. A motivated team has been formed around the tutorial and new people among the former course participants have been involved in the training. The Physics Analysis Groups have produced example code using the common tools and the use of the tools has been made easy.

After the experience of the two years of the training, we can summarize the keys to success as follows:

- people – while it is mandatory to have excellent software experts in the training team, the pedagogical skills and good will are even more important.
- organization – a well established project work-flow makes setting up a new course and sharing the responsibilities easy.
- material – a missing technical detail in the instructions may be overlooked by experts but it may become a show-stopper for the course participants, therefore the exercise material must be checked in detail by non-experts before each round of the course.
- participation – it must be made clear to the participants that the course will take time and they need to work through the exercises – no learning without doing
- feedback – each participant is expected to give feedback and the training team studies the feedback and makes the necessary modifications.

References
[1] CMS Collaboration; The CMS experiment at the CERN LHC, JINST 3 (2008) S08004.
[2] B. Hegner: A Tour through the CMS Analysis Model, in these proceedings
[3] Y. Kuessel (2010): PAT: the CMS Physics Analysis Toolkit, Proceedings, PIC
[4] Merrill, M. D. (2002): First principles of instruction. Educational Technology Research and Development, 50(3), 43-59
[5] Malik S, Lassila-Perini K: A perspective of User Support for the CMS Experiment, in these proceedings
[6] http://espace.cern.ch