SUMMARY OF THE JACoW TEAM MEETING
Thoiry, France, February 2002

J. Poole, CERN, Geneva, Switzerland

Abstract

At the meetings of the Particle Accelerator Conference Coordination Committee and the organising committees of PAC, EPAC and ICALEPCS in 2001 it was agreed that JACoW should operate under the terms proposed at the JACoW Team Meeting in Frascati. In particular, it was greed that the JACoW team should meet at least once per year in order to review technical aspects of the team's work and to define the strategy for the immediate future.

The meeting had initially been scheduled to be in Berkeley at the beginning of December 2001 but it was postponed following the tragic events of September 11th. The meeting was eventually re-scheduled with a somewhat reduced programme to take place in February 2002 at the Holiday Inn in Thoiry, close to CERN. A total of 14 people from 10 institutes attended the meeting and this note summarises the meeting and presents its conclusions. Copies of the slides from the talks are available on the JACoW website [1].

1 PROGRAMME

The aim of the meeting was to review JACoW activities and status and to plan the activities in 2002. As usual, the meeting was based on a number of sessions built around presentations from experts with >50% of the total time allocated for discussion.

The first day was essentially devoted to an in-depth review of PAC'01 proceedings production. Several new techniques were used in Chicago including new templates, a special font set for use in abstract submission and a fully integrated Oracle database. These new features were presented in detail and their performance critically reviewed. The second day focused on the future of JACoW activities and the website. Gerry Jackson, PAC'01 Programme Chairman has initiated a project to archive early editions of PAC in PDF format and the initial results and status were presented. It is intended that these conferences should also be archived on JACoW. The possibilities of including slides, videos and posters in the electronic proceedings was discussed. In the final presentations new software, status of the mirror sites and search engines and the possible use of XML were reviewed.

2 OVERVIEW OF PAC2001 PROCEEDINGS PRODUCTION

Sara Webber presented an overview of the proceedings production in Chicago where a total of 1328 papers were published. Although there was an apparent delay in publication, quality assurance had been completed for all submitted papers by the end of August 2001 and there was then a long wait for submission from eminent invited speakers. The timescale was therefore determined more by political issues than technical matters. By the time of this meeting the proceedings had been published on CD, JACoW, paper and in SPIRES.

There were a number of innovations for the conference including the introduction of compulsory FTP submission for the papers. This resulted in considerable gains in efficiency of the editorial staff but did require additional support staff in the Email/authoring facility. An account system was implemented for authors and this proved useful for contacting authors as well. For their abstract submission authors were permitted to use special characters (a feature previously available when submitting Word files) through a special font which was created for this purpose.

New Word templates were developed for the conference and these have now been adopted as the JACoW standard. They were based on the previous JACoW templates but new styles and macros were incorporated. The editors believe that these had a major impact on the (editorial) quality of papers submitted.

An equivalent of 11 people full time (FTE) were occupied with editing papers at the conference whilst a further 5 FTE's were used for paper acceptance, dotting etc. Support staff for the informatics systems totalled another 7 FTE although there was a peak in the first few days in order to cope with support for FTP submissions. Around 70 computers were installed together with 7 printers and 25 laptop connections.

An Oracle database was implemented for the conference administration and more functionality was implemented within the database than at any previous conference. Processing of papers was tracked through the database and this made it simpler for the staff in reception to check the status of papers and to handle author feedback.

A new script was developed by Leif Liljeby for the processing of the quality assured PDF files. This script handled hidden fields, cropping, page numbering and the conference stamp. It avoided the step of going back to postscript which had been used at previous conferences. The final paper version was prepared using BatchPrintPDF which enabled all of the individual files to be printed, rather than making one large PDF file per volume and the proceedings were then produced from the camera-ready copy.

Some of the usual problems were encountered again, like the authors being unable to follow instructions and this was particularly noticed with the use of the special characters.

[1] http://cern.ch/JACoW/CERN_TM-2002/Talks.html

There were also some difficulties with the creation of accounts for submission, which the submitting authors had to do themselves. The account method was new and it was noted that it should be made clear to authors that it is necessary to have one account per submitting author. In cases where a secretary was submitting the information it is necessary to have one account for each of the primary authors so that the system knows who to contact at the conference in case of problems. This is because each account has only one Email address associated with it.

At a certain point the file upload facility became very slow and it was decided to move to the alternative solution - FTP submission. This is discussed in more detail below. The FTP method sometimes failed because authors only completed one of the two steps required - sending the meta data with the web form and then sending the files by FTP.

Once again the major problem for editors was in dealing with large graphics files. This topic was discussed again later in the meeting - see Section 6. It was also noted that the way in which the quality assurance checks were made involved a sequence of operations which was quite long and that splitting it into smaller pieces would have been more efficient.

Equations in Word sometimes did not reproduce properly if the file was reworked by an editor. This problem was spotted very early and and fixed by installing the Math Type plugin which is available in the public domain. It was explained that authors have found this plugin gives better results than Microsoft's equation editor but that it uses some special fonts/characters which are not available with the basic software from Microsoft. When the paper was reworked, the drivers could not find the characters and therefore the equation was not reproduced correctly.

Sara underlined again the importance of adhering to the ISO9660 standard when naming files and directories. From the beginning of the process, all names should be in uppercase and contain no more than 8 characters and have a three character extension.

3 DATABASE AND ASSOCIATED INTERFACES

Matt Arena was responsible for the implementation and operation of the database and he presented the work which he had done. The installation was made using a substantial PC incorporating 2 18 GByte RAID disks and Oracle Enterprise Edition 8.0.5 on an NT server. A separate Web server was implemented using a 650 MHz Pentium III with 10 Gbyte RAID disk, 1 GByte RAM and running Oracle Application Server (4.0.8.0.2a) on Windows NT4. The file server was used to house the files sent by FTP as well as the PDF files and there was a shared area used by the editors. A total of around 10 GBytes was allocated for the file services.

Nearly 5500 authors were registered in the database and a total of 7000 files were submitted. The system was backed up and exported weekly and there were also daily incremental backups made. The redo log files were also copied to the file server for quick access to enable a rapid recovery in case of a catastrophe. Fortunately no data was lost and no recovery was necessary.

Each editor interacted with the system through web based screens which allowed the execution and management of the editing process. Having such a facility allowed on-line generation of statistics and diagnostics which were also available through Web screens.

The only real problem with the system was related to the web upload of papers. It worked very well during the early days of submission but the huge number of submissions arriving at the last minute caused problems. Matt explained that he thought that the problems were probably related to the fact that although Oracle can handle the transaction rate, the data was all being written to a single disk (tables on one and indexes on another) and therefore the system was limited by the capacity of a single disk drive. He suggested that things would be better if one could use Oracle 9i Application Server together with Oracle 8i or 9i RDBMS. Furthermore if the database was stored across multiple disks using striping, the performance should be considerably improved.

The servers remained at Fermilab throughout the conference and the conference was connected via a T1 link which went direct from the hotel to Fermilab. The backup solution via ISDN was not used.

A large number of utilities were build around and driven by the database, taking full advantage of the possibilities. Many of these facilities such as on-line statistics and multiple Emails (recipients selected according to the status of some data in the database)were created during and after the conference.

4 TEMPLATES AND SPECIAL CHARACTERS

The new templates for Word developed by Sara Webber were offered to authors from the PAC2001 site with a link from the JACoW which explained that the templates were under test. These files represented the first real templates which made use of style files *(.dot)* and included macros. A new set of instructions was also prepared to help authors to use these files and both *.doc* and *.dot* were available. Following the success of the templates they have been adopted as the JACoW standard.

Sara proposed some further changes to the templates following the experience gained and these will be incorporated in the templates for 2002. In addition she made some suggestions for further developments concerning figure insertion with macros and changes for equation and caption styles. The changes which have already been developed included:

- Remove empty carriage return after title
- Changed leading space around table and figure captions

- Changed the insertion method for figure 2 so that the text flow was more logical (following the standard columns)
- added 'body text with no indent' style

PAC2001 allowed authors to use special characters in titles, author lists and abstracts. The authors were able to enter the information using a markup language, similar to LaTeX. This information was translated into unicode or a *gif* representation for the web or a font character number for printed material. A translation was also provided to plain text for use in SPIRES and PDF hidden fields. This facility required that a new font was developed and it consumed a lot of time. Further details about the font are given in the Post Mortem report [1]. Usage of the characters was very limited in titles and author names ($<$1%) but around 10% of the abstracts used them.

In spite of the efforts of the PAC editors, many authors still inserted special characters by cutting and pasting from word which, of course, does not work. There was not a consistent use of special characters by authors, sometimes using the anglicised spelling and others using the font.

5 FONT AND GRAPHICS RENDERING PROBLEMS

In an addition to the original programme, Martin Comyn presented some of his recent findings concerning problems with Adobe Acrobat PDF. The first thing concerned subsetting of fonts when using LaTeX/dvips and subsequently combining several PDF files together. If there are characters used in later documents which were not used in the first document (i.e. not in the font subset of the first document) then these letters may not be printed. The solution to this problem is to turn off sub-setting in dvips by using the *-j0* switch.

Secondly martin reported that he had found that certain versions of Acrobat 4 display but do not print math minus signs ($-$) and uppercase gamma (Γ) symbols. He did not exclude that there may be other characters which have problems but these were the only ones which he had come across. The following table summarises his results:

Mac Acrobat 4	$\times$
Linux acroread4	$\times$
Mac acrobat Reader 4.0.5	OK
Digital Unix acroread4	OK
Acrobat 3 and 5	OK

Acrobat 3 and 5 have not been extensively tested but they seem OK and PC versions of Acrobat have not been fully studied. The investigations are on-going in an attempt to understand the cause of the problem.

The rendering of graphics has been found to vary across versions 3, 4 and 5 of Adobe Acrobat. With 3 there were no problems but with 4, thin lines appeared fainter on the screen and could also disappear completely when printing on a 600dpi printer. When viewing the same file with Acrobat 5 the results were quite different again with heavier rendering on the screen and much thicker lines when printing. One way to get around the problem is to ensure that all linewidths correspond to $>$1.5 pixels in their final form (after scaling).

Finally, Martin pointed out that when printing from a PDF using a Xerox DocuTech system (medium volume printing) an additional layer of software has the effect of reducing linewidths and rendering grey scale translations of colour drawings as if it was printing at 200 dpi.

6 PAPER PROCESSING

Sara Webber described the editorial process developed for PAC2001. At the conference (and during the pre-conference processing) editors were proposed a paper by the database application according to their selection criteria (LaTeX, PC, Mac etc.). A copy of the files was then placed on the desktop for processing and following successful distilling, cropping etc. the PDF was placed in the correct folder (drag and drop) and automatically posted on the web. Papers were assigned dots in the database and the printed copy was marked appropriately so that the dots could be posted on the board. Files which failed to process successfully were copied to the shared area for fixing or to await re-submission.

Most of the normal problems were encountered (bad margins, bad fonts, LaTeX Type 3 fonts) but the most difficult was the large figure problem. Typically an author has produced, or been given, a plot resulting from a tracking program or simulation which produces millions of points or vectors which superimpose. The result is something which takes an extremely long time to display and also makes the final PDF very large. In the discussion in Thoiry it was agreed that JACoW should try to give more help to authors who have difficulties in this area and also to provide some diagnostics for authors who may not realise that they have problems. Tips for authors should include recommended formats for graphics (e.g. *png* or *gif* for line art rather than *jpg* and suggested software for production of bitmap images of huge files (Gemini, Adobe Illustrator, Corel Photoshop and ImageMagic on UNIX).

Sara pointed out that if cropping had been done using the Acrobat feature it was possible that the final result from Leif Liljeby's script for finishing the files (page number etc.) could result in considerable offsets in the margins. If cropping is done using PitStop's page resize feature, the results were consistent - this was therefore recommended as the default method for future conferences. Another margin problem which was quite common at PAC2001 resulted from papers fixed at the conference when there was a mismatch between the paper format of the original document and the paper size/driver setup used when making a new PostScript file. In order to avoid this kind of problem warnings and clear instructions need to be given to the authors. The installation of generic Adobe postscript drivers (avail-

able free from the Adobe website) would also be a help.

It was noted that special care should be taken when re-working Word files (problems like Math Type described above) and the quality control also needs to be made with particular attention to detail. Another Word-specific problem concerns the use of symbols - authors need to select the *Symbol* font when using the **Insert, Symbol** tool. LaTeX specific problems relate to the Type 3 fonts but also the use of GUI interfaces to LaTeX. In general these interfaces like Scientific Word and TeXtures do not give a reliable result and the files are very difficult to re-work.

7 FINAL PREPARATIONS FOR PAC2001 PROCEEDINGS

Sara Webber presented the final stages in the proceedings preparation through quality control to file creation and making the published versions. She explained that the multiple stage quality assurance process was prone to making errors because the controllers could easily loose their place and overlook a step. Creation of the indexes was a straight forward process using scripts and the information in the database. The electronic version of the proceedings was built for viewing with a web browser and therefore the files were HTML based.

In order to number the pages and fill the hidden fields a new procedure was created by Leif Liljeby. This is described in more detail in the Post Mortem paper [1]. This procedure used a program called WinBatch and it was able to perform the following on each PDF file:

- crop the pages
- enter hidden field information
- add page numbers
- add conference stamp and copyright information
- save the output in a new folder

In order for this new procedure to work it is necessary to have WinBatch, Adobe Acrobat 4.x, PitStop and Impress Pro, the latter being plugins for Acrobat. The thumbnails and opening parameters for the PDF files were set using the Acrobat batch processor. Although the proceedings were set up for viewing with a web browser, an index was created using Acrobat Catalog and included on the CD-ROM.

The full suite of Acrobat software for the Reader with search capability was loaded on the CD-ROM. Sara pointed out however that to do this for Acrobat 5 will be more expensive and will require that the publisher (conference) registers with Adobe. The publisher will be required to provide updates of the Adobe software as necessary and also they will have to sign an indemnity clause. It was concluded in the discussion that future conferences will not need to supply the Acrobat reader software because it is now very widely distributed and people will not need to get it from the CD-ROM.

8 SCANNING OF PAC PROCEEDINGS FROM THE PRE-ELECTRONIC ERA

Gerry Jackson presented the status of the project and described the performance obtained so far. A description of the project can be found in the write-up which has been published since the team meeting [2]. The project concerns the conferences held from 1965 to 1993 – a total of 15 conferences, 32 volumes and nearly 27000 pages. To date about half of the funding required for the whole project has been found and efforts to find more continue. The process, which is performed by a private company, involves scanning each page, performing character recognition (OCR) and making a PDF file per paper.

A test has been made on one volume and the results have been very encouraging. Several issues were addressed in the test:

- PDF File size vs. quality
- Indexing for full text search
- Generating the data for the hidden fields
- Inserting data into the hidden fields and adding conference stamp etc.

The quality of text is not really an issue, the problems are more associated with the graphics. Scanning at a higher density improves the quality but at the cost of file size. The OCR works well although it cannot handle maths or special characters. The result of the OCR sufficiently good that the full text search in the final PDF works efficiently.

In the next stage it is proposed to optimise the scanning parameters (contrast, density) and to develop algorithms to extract the data for the hidden fields. Even at 200 dpi scanning density ($\Rightarrow$ 320 kByte per paper) it would still be possible to fit one conference on a CD-ROM.

9 SLIDES, VIDEO AND POSTER INCLUSION

John Poole presented an analysis of the proposal to include slides, video and posters in the electronic publication. Since being asked to investigate these possibilities PAC2001 has implemented slides and video in their proceedings. It was agreed during the discussion that posters are not well suited to view on a PC screen. Furthermore the variety of software packages used in poster preparation would make the task of the editors very complicated. As a result it was agreed that for the time being no effort will be devoted in this direction.

PAC2001 linked slide presentations with streaming video for the opening and closing plenary sessions. There are links from the Table of Contents in the proceedings to the FNAL server which delivers both the video and slides to a web browser. This system works very well and was an impressive development. John pointed out that, given the extra resources required during and after the conference, including slides and video is a very worthwhile exercise. It was underlined that it is essential for the smooth

running of the conference sessions that electronic presentations are loaded on to conference computers before the session. This allows verification that the slide show will work and at the same time the author's electronic files are captured and available for future use. Gerry pointed out that all authors had been instructed to come to the podium with transparencies as well so that any electronic hiccups would not delay the session.

Based on experience from the Chamonix performance workshops it was concluded that the extra space required on the CD-ROM for PDF versions of transparencies would be around 1 MByte per talk. Since it has already been noted that there will only be room for one conference, there should be plenty of room for the transparencies.

Following the discussion it was agreed that EPAC2002 team would publish as many of the oral presentations as possible. Electronic presentations will be encouraged but slides will be scanned if there is no alternative. The additional resources to do this work were estimated at 2 full-time-equivalents during the conference and about one man-month after the conference.

10 NEW PRODUCTS

Leif Liljeby presented a summary of his evaluation of some new products. He explained that Acrobat 5 does not bring any particular improvements but the plug-ins are more interesting. He explained that PitStop remains a very useful tool and he recommended that it is used for making global changes to files through action lists. He also explained that for the next conference it will be interesting to continue using Impress Pro for putting the page numbers and conference stamp into the files. A useful tool for editors is the Gemini plug-in which can be used for exporting large graphics so that they can be converted to smaller bitmapped images for re-insertion.

11 XML AND JACOW

Ivan Andrian reported on the possible uses of XML by JACoW. It was already noted at the JACoW Workshop in 1999 that this was an interesting development but that it was not yet mature enough to be of use for us. It became clear that the technology has developed rapidly in the last couple of years and although it is not yet at the stage where we can use it, the next EPAC (2004) will probably see its introduction.

The technology has its strength in the fact that it separates the scientific matter from the way in which it is presented. This means that authors can concentrate on the content and it will be easier for the editors to control the presentation. Because it is a markup language it will simplify data gathering for the conference database and styles can be imposed. It will also facilitate much greater portability of the information and facilitate electronic publication.

Given so many advantages and the fact that it will soon be in common usage, it was agreed that a pilot project will be launched within JACoW in order to learn about the technology and ultimately to develop the necessary templates for EPAC2004.

12 SEARCH ENGINES

Pascal Le Roux presented an analysis of the use for search engines. He explained how the search engine installed at CERN works. The system is based on a custom interface to an Inktomi Enterprise Search 4.2 engine running on a Windows machine. The engine is set up to keep an index of the whole CERN intranet which corresponds to about one million documents. There is a programmable limit on the size of files which are indexed which is currently set around 4 MByte. An interesting feature of this particular search engine is that it indexes both the full text in a PDF document and the meta data (hidden fields) and this is why JACoW adopted it. The system has proved to be very reliable and although it is quite expensive, CERN was able to negotiate a very interesting price. Currently there are around 1000 JACoW searches per month.

Pascal explained that there are many other products available which could also do the job but they are of similar cost or more expensive. Another way of attacking the problem would be to use a remote engine like Google which would be free, but without the same rapid indexing of new files (about one month rather than a few days) and without the full functionality (no meta data capability). It was concluded that the engine at CERN is perfectly adequate for JACoW and has the advantage that it is supported centrally. If it was necessary to look for search engines for the other mirror sites, Pascal pointed out that FNAL has an Inktomi installation.

13 MIRROR SITES

Martin Comyn reported on a series of tests which he had made concerning the performance of the JACoW sites. He explained that the full set of conferences was not available from the ANL website because they did not believe that all of the copyright issues had been resolved. It was the opinion around the table that this had, in fact, been resolved after the discussions concerning PAC'97.

Martin made a series of tests at all times during a 24 hour period and found that rate of delivery of files (kBytes.s^{-1}) was fairly constant and that from Triumf, the ANL site was about twice as fast. However, the delivery time of an average sized file from CERN was still only $\sim$3 s. Martin also asked Yong Ho Chin to make some tests and he found that CERN was about 30% slower than from Triumf and ANL was about a factor of two slower than from Triumf. Martin pointed out that Yong Ho was in the process of implementing a mirror site in Asia and that it should be available before the end of the month (subsequently confirmed and is available via the JACoW site at CERN).

Martin reported that in his discussions with Bill McDowell, the ANL JACoW webmaster, he had learnt the ANL

were thinking of purchasing an Inktomi engine for their site. Unlike the ANL site which uses Wget (GNU software), KEK were using windows software but they are not planning to install a search engine.

Martin concluded by suggesting that the issues of copyrights should be cleared up as soon as possible and that if there were problems concerning the purchase of search engines, then some funding should be requested from the conference series. He also pointed out that as soon as the system is fully functional we should be more pro-active in publicising the existence and features of JACoW.

In the discussion it was concluded that as far as North America is concerned, the mirror site does not seem to be really necessary. The situation was less clear for Asia where it was felt that the internet connections to some countries and institutes may have less bandwidth and therefore "local" copies might perform better.

14 FUTURE IMPROVEMENTS

During the discussions throughout the meeting a number of suggestions were made concerning possible improvements in our activities and these are reported here.

- It was suggested at an earlier meeting that Adobe PostScript drivers could be used for the production of files for distilling on non-Unix platforms. These drivers are available free of charge from the Adobe website.
- There was a lot of discussion about the difficulty to obtain author's details for the conference databases. Problems such as authors spelling their names differently (quite common when it is a translation from Cyrillic) or writing their institute differently cause a huge time loss for conference organisers. In order to improve this situation it was suggested to use the person's Email as a unique identifier. Although it is recognised that people do move around the vast majority will keep the same details from one conference to another.
- It is customary now (at least for PAC and EPAC) for authors to have an account which they use for submission of their contributions. Often when a secretary has submitted for a group of people there have been some difficulties. It was felt that some additional education was required in order to make it clear that authors need to have one account per submitting author.
- The main problem in paper processing is now large files and it was agreed that authors need help in this domain. Some information will be prepared for the web which will tell authors how to check and see if their paper may have problems and it will propose some solutions.
- In the future no Acrobat software will be installed on the CD's – this will simplify the work of the editors and liberate some more space.
- Michael Böge mentioned that an increasing number of people are converting the HTML abstract brochure to a format suitable for downloading to a Palm Pilot. It was agreed that a suitable file be made available on the website so that people can just download the file directly.
- A JACoW pilot project aimed at exploring the capabilities of XML will be launched this year.
- The service to authors where they can submit their paper and receive a PDF version in return to check its performance should be re-introduced with some enhancements. It is proposed to introduce cropping and showing the text box on the finished PDF for the authors - in this way they can check out their margins and quality of paper.
- A new service for LaTeX authors should be introduced similar to that provided by LANL. In this way LaTeX processing and PostScript production could be made in an environment which will ensure that there are no Type 3 fonts (if the author uses the system!).

15 NEXT MEETING

The next meeting of the team should be held towards the end of 2002. It will be held in North America, possibly in the Chicago or Oak Ridge areas.

16 ACKNOWLEDGEMENT

I would like to express my thanks to Christine Petit Jean Genaz for organising the meeting and all of the speakers and the participants for their efforts which made the meeting such a success.

17 CONCLUSIONS

Once again bringing the team together has facilitated improvements in our facilities and has greatly helped the 2002 conference organisers. The success and usefulness of JACoW is underlined by the increasing support being received from the parent conferences and we can look forward to an expanded and improved website in the coming year.

18 REFERENCES

[1] S. Webber, P. Lucas and M. Arena "Post Mortem of the Electronic Publication of the PAC 2001 Proceedings", published by JACoW, http://cern.ch/JACoW/organisers/docs/PAC01-PM.pdf, February 2002.

[2] G. Jackson, Archival of PAC Proceedings from the Pre-electronic era, March 2002. ... *waiting for correct reference*