EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN – AB DEPARTMENT

AB-Note-2005-016

POST MORTEM OF THE ELECTRONIC PUBLICATION OF THE EPAC 2004 PROCEEDINGS

J. Poole, CERN, Geneva, Switzerland

Abstract

The proceedings EPAC'04 were the fifth in the series to be published electronically. This report describes the preparations before the conference, the activities at the conference and the work afterwards which was required to produce the CD-ROM and HTML versions of the proceedings. The process for this conference was based on the JACoW Scientific Programme Management System (SPMS) which was in full production for the first time. The conference was very well attended and even more papers were published than before. This report is the fifth in the series of post mortems and reflects what was new in 2004.

Geneva, Switzerland

March 10, 2005

Contents

1 Introduction

This report describes the new features in 2004, analyses the performance throughout the procedure and presents the relevant statistics. The topics covered include the preparations before the conference, the activities at the conference and main features of the JACoW Scientific Programme Management System (SPMS) used for the organisation and management of the scientific programme and for the processing of contributions. In a departure from previous practise, no paper volumes were produced.

Major progress for EPAC'04 was the introduction of the SPMS, the underlying software tools and techniques were very little changed with respect to earlier conferences. The Lucerne conference was even bigger than Paris with 937 papers published but requiring less manpower than in 2002, if one discounts the development effort of the SPMS. The final version of the proceedings was available on WWW in just over eight weeks.

This report does not repeat issues discussed in previous post mortems and readers are therefore recommended to refer to the earlier ones [**?**] in order to obtain a full picture.

2 JACoW SPMS

In recent years JACoW collaboration conferences have developed software for handling abstract and paper processing and at the JACoW collaboration meeting following the last EPAC conference it was agreed to develop a database system for for the whole process of scientific programme management.

The system is built around a central repository of author profiles and institute data. Each conference has its own database in which the scientific programme is defined, abstracts are stored and meta data for papers are stored. Files submitted by authors are uploaded to a separate file server and at the same time, the meta data is sent to the conference database. Interaction with the database for users, editors and administrators is via a web interface.

EPAC'04 was the first conference to use this system on-line although BIW'04, Cyclotrons'04, LINAC'04 and FEL'04 used the system in stand-alone mode i.e. with no connection to the central repository. The advantage of the system is that the quality of data available from the central repository is better than anything that has been available before and in principle, authors only have to enter their details on the first occasion when they use the system. A huge effort was invested in checking and correcting the profile and institute data but this effort did not have to be re-invested for the conference as had been the case for each conference in the past.

Several hundred sets of validated profiles and over 600 institutes were available after abstract submission (at the time of writing 3100 of the 7400 profiles have been validated and there are 1100 affiliations). At the time of writing, the number of personal profiles has more than doubled because the authors for the PAC'05 conference have created their accounts.

3 Conference Website

A domain name (EPAC04.ch) was purchased in 2002 and the conference website was set up on a server at PSI. The web pages were built using a proprietary tool (GoLive) and the result was very satisfying aesthetically. The web pages were editable from CERN, ETH and PSI. Creating and maintaining pages was rather difficult because navigation was not very transparent, however, this is not an intrinsic problem with the system. The titles of all pages (which are in the HTML meta-data) appear in the browsers title bars but these were not meaningful to users or web-authors and this detail could be improved in future.

4 Abstract Submission

Abstracts were uploaded directly to the database and because the author data had been validated in the repository, it was possible to produce the basic information for the Abstracts Brochure automatically. The associated Author Index was prepared in a few minutes – something which had previously been a tedious task.

The SPMS enables statistics to be extracted very easily and an example is shown in Figure 1 where the distribution of abstract submissions in time is shown. This plot is no surprise to any experienced organiser because it clearly shows that authors always wait until the last moment before submitting - 50% of abstracts arriving on the last two working days at the deadline. The dip between the two peaks corresponds to the weekend which preceded the Monday deadline.

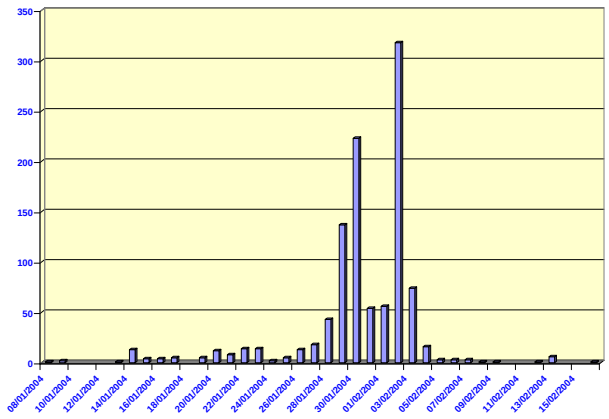

Figure 1: Number of Abstracts Submitted per day

5 Paper Submission

Templates for the preparation of papers were unchanged from those used for PAC'03. Paper upload was to a file server which was mounted on the PSI afs system. The

deadline for paper submission was Wednesday June 30th at 24:00 CET, which allowed a few days for pre-conference processing.

6 Paper Processing

One of the key factors in the success of the computing facilities for EPAC'04 was the meticulous planning and testing carried out by the LOC. The IT services were set up well ahead of the conference and the networking facilities were tested at the conference site. A complete installation of an editor's computer was prepared many weeks ahead of the conference. After testing and modifications to the editors' software systems, a master image of the system disk was prepared and this was subsequently installed on the production machines.

The procedures for processing were much the same as in previous years apart from the use of the SPMS for management of the process. The SPMS interface delivered papers to the editors, kept track of their activities and comments and uploaded the processed files to the file server. Editors were encouraged to record clear explanations of any problems encountered so that the staff in the paper reception office would be able to explain the problem to the authors.

6.1 Pre-conference Processing

We were fortunate to be able to carry out pre-conference processing at the conference venue, albeit in a different room from the definitive office. The full system was set up for eight editors with 8 PCs and one Mac and processing started on the Thursday before the conference. The deadline for submission of papers was midnight CET on the Wednesday and we were surprised to find that more than 850 papers had been submitted by the deadline (∼90% of the total). After a short period of editor training, processing was able to start.

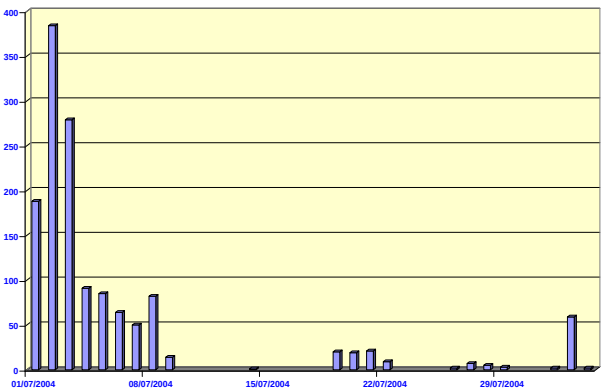

Figure 2: Number of Papers Processed per day for EPAC'04

The numbers of papers processed per day is shown in Figure 2. The statistics support the analysis made in 2002, which revealed that an editor can process around 35 papers per day if all is working well. Things went so smoothly in fact that it was necessary to stop editors processing so that there would be some papers available for processing during the conference week – an important feature for the training of future editors.

6.2 During the Conference

The transition from pre-conference processing was very smooth as the conference office was set up on the Saturday whilst pre-conference processing was continuing in the other office. It was therefore possible to move seamlessly into the conference office on the Sunday.

There were not many papers which required processing during the conference and it was possible to move on to Quality Assurance (QA) as had been planned. Work continued smoothly and all papers had been processed and QA'ed by Thursday evening of the conference week. There were of course, a number of problem papers outstanding and we were still waiting for a few re-submissions.

6.3 Post Conference Activities

Post conference activities were mainly concerned with the cross-check of Titles and Author lists which appeared on the paper with what the authors had entered in the database. There were many discrepancies in this area because authors had either not bothered to enter the names of co-authors or they had created duplicate profiles for them.

There were probably more re-submissions to make small changes than in previous years, but the number was not excessive. The problem papers were all solvable either by the editors or by the authors themselves.

The speakers slides were uploaded in the same way as the papers and then transferred to the conference computers. A backup copy was made of the files on the conference computers and this was used to prepare the files for publication. All files were converted to PDF using the PDFMaker plug-in for PowerPoint files and by distilling the LaTeX sources. This process took some time because several attempts were necessary for some files which did not perform very well. About one man-week was consumed in this activity.

The final stage, the preparation of the files for the website and CD, was done by Volker Schaa at GSI. The final files were compressed and packed in 'tar' file which was placed on a webserver for download and unpacking at CERN.

6.4 IT Resources

The hardware requirements for 2004 were based on EPAC'02 and PAC'03 experience where about 10% of the papers were prepared on Macintosh. There were two Macintosh and 16 PC's in the processing office and a further 4 PC's in the paper reception office. The computers were linked by a local intranet. The printing service was implemented using 2 colour and 2 monochrome printers, shared between the two offices.

10 Gbyte of disk space were allocated for the conference and a further 10 Gbytes were reserved in case they were needed, but only 8.2 Gbytes were required. This included the multiple versions of files and all of the talks.

The LAN at the conference centre was connected to SWITCH (the Swiss Education and Research Network) via a third party link which offered a bandwidth of 35 Mbit/s. PSI supplied the switches, routers (WLAN) and firewalls and the full time support person who was available at the conference site throughout our presence there (from pre-conference to dismantling).

The software inventory was as follows:

PC

Windows XP Professional
Microsoft Office Pro
Internet Explorer 6.0
Netscape Communicator 7.1
Mozilla 1.6
WinZip 9.0
Adobe Acrobat 5.0.5
Adobe Distiller 5.0
Enfocus PitStop Professional 6.0
WinEdt 5.3 30-day trial version
GSview 4.3 (Ghostgum Software Ltd)
MiKTeX 2.4
LateX2e 1.6
WS_FTP Pro 8.03
WinSCP 3.6.1
PuTTY 0.53b
WRQ Reflection X 10.0.0
McAfee Virus Scan 4.5.1
Agnitum Outpost Personal Firewall 2.1 30-day trial version

Equivalent software was installed on the Macintosh machines.

As usual there was strong support from the JACoW collaboration with support from CYCLOTRONS, FEL, ICALEPCS, LINAC, PAC and RuPAC for the proceedings and processing offices. In all there were about 16 people assigned to the processing office.

7 Publication

The efficiency of processing at the conference and the flexibility from the SPMS meant that it was possible to publish the pre-print version of the proceedings on the web just 9 days after the conference.

The mechanism to recover the files from PSI to CERN was a version of the upload script which used an input file containing the list of files which had the 'OK_to_Publish' flag set to 'Yes'. The files were downloaded from the PSI server to a computer at CERN and then they were copied to the webserver. The pre-print version used index and contents files generated dynamically from the database using the EPAC'02 script to link to the processed PDF files. It was necessary to use the EPAC'02 scripts because a problem was discovered with the SPMS installation at CERN (see Section 10.8 below).

The final version of the paper PDF files and web pages was transferred from GSI and copied to the JACoW server. These files were also copied to a CD master which was sent to PSI for final testing and production.

8 Statistics

8.1 Manpower

The manpower used in 2004 was much more than in 2002 but it includes the development of the SPMS, a one-off task estimated to have required ∼24 man-weeks. The benefits in the long term are obvious but even at the conference there were significant advantages. The distribution of manpower resources involved in the proceedings is given in Table 1.

Table 1: Manpower Resources for EPAC Proceedings in Man-weeks

	2002	**2004**
R & D	4	26
Planning	2	2
Build/maintain WWW pages	1	1
Author documentation	1	0
Server setup	2	2
Abstract Processing	5	0
Infrastructure	2	2
Pre-conference Processing	3	6
Processing at the Conference	15	13
Post Conf - local	4	2
Post Conf - CERN	14	8
Total	50	66

8.2 Computer Platforms

There was a further slight shift away from Macintosh platforms (22% in 98 and 12% in 2000, 10% in 2002) with only 6% in 2004 and the number of Unix users remained about the same.

8.3 Software used by Authors

The distribution of software packages used by authors remains dominated by Microsoft Word and LaTeX (see Figure 3). The percentage of LaTeX users has remained constant whilst non-preferred software is restricted to two papers – from the same author who is using FrameMaker. This author received red dots for his papers again (we have had problems with him at all of our conferences) and had to re-submit.

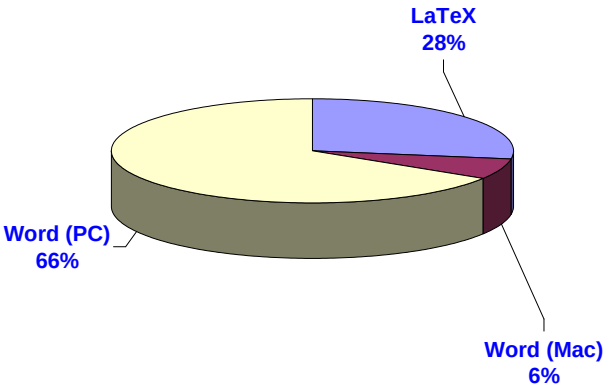

Figure 3: Software used for the preparation of EPAC'04 papers

8.4 Failure Rates

The overall quality of papers submitted continues to improve and editors are able to apply higher standards each year. The fraction of papers with real problems was less than 10% in 2004 but the number of papers marked for proof reading increased to 37% (see Figure 4). This is also possible because the software has improved and it is much easier to fix many problems either using PitStop or by reworking the original document. It has been a policy for some years to invite authors to proof read all papers where the original document has been re-used to make a new PDF. This process is necessary because the author's installation may differ from that at the conference and there may be font or plugin problems which are not immediately obvious to the editor.

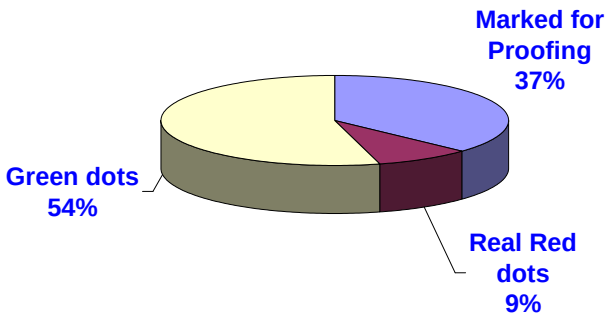

Figure 4: Results of Processing at Lucerne

In Paris 27% of all papers were marked for proof reading but there were only 14% of the papers which had real problems. Figure 4 shows how this has evolved with <10% of real read dots in 2004.

8.5 Fault Analysis

The unidentified problems appear because there is a check box labelled 'other' in the editor's interface. An analysis of the comments in the database shows that in the undefined category there were perhaps 50% of cases where the editor could have attributed the fault as either a format or font problem (or a combination of the two). In the remainder, there were a large number where the editor experienced complications with the processing of the file. In other cases it was simply necessary to change the status of a paper to recall the author (for example to proof read after a request to modify a paper which had no fault). It is therefore probably worth adding a further two fault categories - 'Processing complications' and 'Author recall'.

Table 2: Analysis of problems encountered with papers submitted to EPAC 2004

	Percentage of all papers
Format problems	20
Unidentified problems	9
No PostScript	7
Font problems	5
Type 3 fonts (LaTeX)	2
A4 printed on US paper	2
Unusable files	2
Slow graphics	2

Many of the 'No PostScript' papers were submitted with PDF - this a problem which is getting worse and will require some action for the next conference.

9 Logos and Poster

From an early stage in the organisation of the conference high quality graphics files (conference poster and similar materials) were widely available in many formats and this was highly appreciated by the many actors who need these.

10 Problems Encountered in 2004

There were no significant problems encountered during 2004 and those reported below had no significant impact on the efficiency of the whole process. There are a few things which could be avoided in future which would make the process even more efficient and these are described in the following sections.

10.1 Processing

A total of ten papers slipped through beyond final QA with the wrong paper size. These were detected by a visual basic script which checks the number of pages and the paper size. Four papers had the wrong number of pages in the database. It would probably be useful to use the script to enter the number of pages in the database once the editors have checked the files and removed any blank pages etc.

10.2 Action List

The specification for the PitStop Action List which is used to change the media box (paper size) to JACoW format was incomplete. In addition to changing the media box, it is necessary to remove the crop box. If this is not done, the

perl script which prepares the files for the CD does not put the bannners and page numbers in the correct place on the page and in some cases the script fails. In order to fix this problem the defective files were reprocessed using PitStop Server to run an Action List which removed the crop box. About 10% of the PDF files had to be re-processed in this way to remove crop boxes.

10.3 Dotting Board Format

The dotting board was produced by printing the programme codes on small format sheets of paper (A4) which meant that the sequence of papers was not obvious to authors searching for their paper. It would be clearer to print the codes for one day on one sheet, which should be possible using A2 or larger size paper.

10.4 Acrobat 6

It is JACoW advice to only use Acobat 5 software because some features cause problems during processing and PDF version 1.5 files generated by Acrobat 6 generate error messages when opened with earlier versions of Acrobat. However, it is not possible to purchase Acrobat 5 for MAC OS/X and therefore we had to install the latest version. In principle, if the distiller parameters are set to have PDF v1.2 compatibility, the result should be acceptable. However in spite of using it in this way a large fraction ($>$50%) of the files processed on the MACs ended up as v1.5 and generated errors when opened with the earlier versions of Acrobat. It seems that if there is any modification made to the file in Acrobat 6, then the file is saved as v1.5. A possible remedy for this problem would be to use the optimiser tool which is new in Acrobat 6 and which allows one to save the file with an earlier version compatibility.

In view of the pressure to prepare the files very quickly this problem was not studied in any great detail before fixing the problem. The solution which was implemented involved opening the files in Acrobat 6 on a PC and making a PostScript file which was subsequently distilled using Acrobat 5 on a PC running Windows XP. Given the font issues with Windows XP (reported at the JACoW Team Meeting in 2004) it is likely that the resulting files used fonts which were not available on the PC. It remains a mystery why some files re-processed in this way were still acceptable and others not.

10.5 PitStop License

When the disk image was being prepared for the computers to be used by the editors at the conference, the wrong version of PitStop was installed. A local copy which was a single user version was used in stead of the trial version. An agreement had been negotiated with Enfocus (manufacturers of PitStop) to use the trial version for the duration of the conference in multi-user mode but unfortunately it was not installed. All editors have to use PitStop to change the media box in every file, so a single user license was not appropriate. Some ingenious fixes were applied to recover the situation at the conference centre.

10.6 Missed one paper

When the proceedings were finally published we received a complaint from an author that his paper was missing (although it was many weeks since the pre-print version had been announced). It was found that this paper had failed final QA and subsequently been fixed but the database did not reflect this and therefore the conditions for 'OK_to_Publish' were not satisfied and the paper did not appear in our list. As a result it was necessary to repage number the whole proceedings and remake all of the files for publication. Fortunately this only means running the scripts again and transferring the result back to CERN. A complete cycle takes 4 hours, most of which is the manual interventions on the files. The transfer of 500 Mbytes of data from GSI to CERN, unpacking the files and then uploading to the webserver takes about one hour in total – it only took around 25 minutes to run the scripts.

10.7 Speaker Presentations

There is room for improvement in the way in which speaker's presentation files are named and stored. Speakers were asked to name their talks using the programme code and appending 'talk' to it. This of course resulted in filenames which were longer than 8 characters if the speakers followed the instructions. There were of course a number of files called 'epac-talk' and many other problem cases. The problem was resolved in the end by storing the processed files in a separate directory and using the programme code to identify them. Some improvements in the way this is handled from the uploading stage through to final processing would be beneficial.

Some difficulties were encountered in the preparation of the PDF files from the speaker's PowerPoint presentations. One particularly annoying aspect was that semi-transparent fill areas are not handled well by Acrobat - the resultant PDF is extremely slow to display and does not have a true rendering of the original effect. It was necessary to edit the user's files to remove this type of fill wherever possible. A similar problem was experienced with shadowing and a similar remedy was applied. However, the full range of distiller parameters was not explored and this may offer a better solution for the future.

The InDicCo system was used to manage the videos of the oral presentations. It was necessary to introduce the filenames manually into the database and this was a source of error. In the published version on the web, there were two talks where the URL was incorrect and this was the situation for three weeks before it was spotted. It is interesting to note that this implies there is not much call on the links to the videos.

10.8 Making the CD

A system was developed in the SPMS to create all of the files for the CD using Oracle tools and software. These procedures write the files on the system disk of the machine where Oracle is running. At CERN this capability is

blocked for security reasons and is a hard constraint which cannot be avoided. For this reason EPAC'04 opted to use the scripts developed by Volker Schaa to do this.

It was not realised that the built-in SPMS procedures could not be used before we were ready to produce the final website and therefore there was a period of intense development in which the scripts developed for DIPAC'03 were modified to suit. Data is dumped from the conference database into an XML file which is then used by the scripts. The processed PDF files (formatted, but no hidden fields, banners or page numbers) were downloaded to GSI using WGET and then the scripts were run.

Following a decision at the JACoW team meeting in 2003, authors were not requested to submit keywords. Acrobat is used to extract the text from the PDF files and a perl script analyses the output and counts the number of occurrences of keywords corresponding to the official list. The highest scoring keywords are used.

It was found that there were a number of files where Acrobat was unable to extract the text and these had to be re-processed. For the most part these were files which had been processed on a MAC and which had finished up as PDF version 1.5 (Acrobat 6) but it was not all such files. These files had been re-processed as described in Section 10.4 and it was necessary to produce new PDF files from the original sources. This was achieved by processing them on a PC and therefore the new PDF files were only used to generate the text files for analysis. The underlying problem in this case seemed to be related to the fonts (or rather the lack of font recognition on the PC after the re-processing).

When contributions have not been received for oral presentation there is obviously no paper but there is a video. There were five such cases in 2004. It was necessary to modify the website by hand to introduce 'Contribution not received' on the appropriate web pages (session and classification indices for each paper i.e. ten files to modify). It would be good if this could be handled by the scripts since this is a time consuming operation and is obviously prone to errors.

10.9 Library Data

JACoW has agreed to make library data available as a part of its service and this means preparing files in the appropriate format which contain citation and indexing information. One essential part of this is the keyword data which must therefore be available in the database so that the scripts can extract the information. The built-in scripts were modified so that they could be run interactively and the output data captured in a spool file. As a pre-cursor, however, the keywords which are produced by the scripts have to be uploaded into the database and this was done via Excel and Benthic (proprietary interface to Oracle). It may be interesting to investigate the possibility to extend the functionality of the scripts to prepare the library data in Open Archive Format and for SPIRES.

10.10 Files for the CD

EPAC has stuck with ISO9660 standards for the CDROM and this always proves to be difficult. The filenames which are produced by the database and scripts are easy to keep to the uppercase eight character name and three character extension requirement. However, there are also hundreds of other files like the photographs and their associated web pages which are concerned. In principle these filenames can be changed using scripts, but it is not obvious to ensure that all references (URLs) are changed as well. For this reason, the photos and associated file names were not forced to uppercase, but the names were restricted to <8 characters.

The source CD for the master was prepared by copying the files from the JACoW website which itself had been prepared using a number of tools. During the migration from pre-print version to final version some files had been moved from one location to another using FrontPage which unfortunately changed URLs to maintain links. As a result, although the links were still relative, they pointed to pages in a directory which was in a different tree and which was destined for the CD. This was simple to fix but is the type of problem which can be very hard to spot.

When the first version of the proceedings was in preparation it was found that there was not enough room on one CD. It is clear that technological advances have led to authors including more and more photographs and complex graphics in their papers with the result that the average PDF file size was 425 kbytes compared to 300 kbytes in the year 2000. However, this was not the main problem which was the speaker presentations which had to be reduced to about 60% of their initial size in order to fit them in.

11 Acknowledgements

The very successful IT infrastructure was the result of a great deal of hard work by the following people:

Remo Rickli	Network
Lucas Wacker	Presentation manager
Martin Heiniger	Presentation hardware
Stephan Egli	IT
Xavier Guitierezs	PC Hardware
Laurin Müller	PC Software
Lucas Sekolec	MACs
Jürgen Baschnagel	Printers & PC h/w
Kurt Vollenweider	Printers
Edgar Barabaras	afs
Colin Higgs	Installation at KKL
Jean Renaud	Installation at KKL
Thomas Federer	Installation at KKL

The EPAC editorial board is greatly indebted to the following group of people who worked very hard on the electronic processing of the papers in Lucerne.

Ivan Andrian (Trieste)
Matt Arena (FNAL)
Jan Chrin (PSI)
Cathy Eyberger (ANL)
Frank Gerigk (RAL)
Martin Heiniger (PSI)
Charlie Horak (SNS)
Leif Liljeby (MSL)
Pascal Le Roux (CERN)
Michaela Marx (DESY)
Jeff Patton (SNS)
Volker Schaa (GSI)
Evgenia Shirkova (JINR)
Toshiya Tanabe (RIKEN)

The processing staff could not have completed their task without the support of the other volunteers in the paper reception office:

Mariarita Ferrazza (Frascati)
Sheila Poole (Thoiry)
Pina Possanza (Frascati)
Kathy Rosenbalm (SNS)
Sue Waller (DL)

Finally, I would like to thank Jan Chrin and Detlef Vermeulen, both from PSI and serving on the LOC, for providing such an efficient infrastructure for the proceedings and for the information which I have reproduced in this report.

References

[1] J. Poole, “Post Mortem of the Electronic Publication of the EPAC’96 Proceedings”, CERN-SL-Note 96-68 DI, November 1996.

[2] L. Liljeby and J. Poole, “Post Mortem of the Electronic Publication of the EPAC’98 Proceedings”, CERN-SL-Note 99-015 DI, February 1999.

[3] S. Webber, P. Lucas, M. Arena, “Post Mortem of the Electronic Publication of the PAC 2001 Proceedings”, http://cern.ch/JACoW/organisers/docs/PAC01-PM.pdf, January 2002.

[4] P. Le Roux and J. Poole, “Post Mortem of the Electronic Publication of the EPAC2000 Proceedings”, CERN-SL-Note 2001-007 DI Revision 1, April 2001.

[5] M. Jouvin, “EPAC 2002 Computing Post-Mortem”, http://cern.ch/JACoW/organisers/docs/computer-pm-2002.pdf, August 2002.

[6] P. Le Roux, Ch. Petit-Jean-Genaz and J. Poole, “Post Mortem of the Electronic Publication of the EPAC2002 Proceedings”, CERN-SL-Note 2002-042 MR, December 2002.