REPORT OF THE JACoW WORKSHOP ON DATABASES FOR CONFERENCE PROGRAMMES AND PROCEEDINGS

M. Comyn*, TRIUMF, Vancouver, B.C., Canada
I. Andrian†, Sincrotrone Trieste, Trieste, Italy

Abstract

Brief outline of workshop, dates, location, aims. Motherhood statements about JACoW.

Handwritten but too provocative to publish! Will be toned down over Christmas.

INTRODUCTION

Brief summary of John Poole's ground rules, expectations and dreams.

InDiCo

Awaiting Thomas' paper to prepare a complete summary.

SCOPE AND PRELIMINARY SPECIFICATION FOR A DATABASE FOR CONFERENCE PROGRAMMES AND PROCEEDINGS

This free format session, chaired by Martin Comyn, sought to specify, in broad terms, the essential components of a database to be used for all stages of conference abstract, programme and proceedings management.

Glossary of names

As each JACoW conference series uses a different set of (sometimes conflicting) names for a similar set of committees and functions, a generic list of acronyms and definitions was created during the course of the discussion, as shown in Table 1.

Table 1: Database user acronyms and definitions.

Acronym	Definition
LOC	Local Organizing Committee
OC	(International) Organizing Committee
SC	Scientific (Programme) Committee
SCC	Scientific (Programme) Committee Chair
PE	Proceedings Editor(s) (Board)
POS	Proceedings Office Staff
BUS	Business: Delegate Registration, Finances
DBA	Database Administrator
SYS	Computer Systems/Network Personnel

* comyn@triumf.ca
† ivan.andrian@elettra.trieste.it

Initial specification of database components

Table 2 summarizes the essential database components identified by the workshop delegates, the basic requirements and the various users of each component indicated using the acronyms defined in Table 1. Although delegate registration could be included in the conference database, many conferences hire professional conference organizers to handle these aspects or use lab personnel and existing systems which are not easily integrateable with the database being developed here. It is quite likely that the use of InDiCo for delegate registration could become an obvious option, and then links between the two systems could be developed.

INFORMATION ANALYSIS

Uniquely identifying and tracking authors

There was a general consensus that uniquely identifying and tracking authors through the years as they change affiliations or join multi-institution collaborations, would be extremely valuable. A single database could then be maintained and used by all JACoW conference series, with each sequential conference helping to keep the data current.

Creation and maintenance of profiles

The initial author database could be created by amalgamating and rationalizing the existing EPAC, PAC and other JACoW conference author databases. Authors would then be invited to access their profile and update it wherever necessary. This procedure, although initially quite arduous, would result in a far higher degree of standardization regarding affiliation names, addresses and contact information. In future, all authors would have to be registered in the database before they could be included on abstract author lists.

Finding a profile A simple interface would have to be developed to allow authors to find their own profile, or correctly identify those of co-authors. A simple Web search form would prompt users for the following sequential information, and present a list of possible matches at each step until the correct author is uniquely identified:

- last name;
- first name;
- middle name;
- e-mail address;
- affiliation.

Table 2: Database functions, requirements and users.

Function	Comments	Users
Delegate registration Use InDiCo?	PE need data for cross checks Unique key for people? Problems: multiple affiliations/addresses	BUS, LOC, SCC, PE
Author database	All authors including co-authors Create author profiles Problems: multiple affiliations/addresses	PE, SC, SC, LOC, PE
Abstract submission	Input against existing author profiles All via WWW All posted on WWW	SCC, SC, PE, DBA, LOC
Programme committee	Determine whether invited/oral/poster presentation	SCC, PE, DBA, LOC
Conference session ordering	Determine session/time/place order Assign order and paper ID# as late as possible to exclude withdrawn abstracts	SCC, PE, DEA, LOC
Abstracts brochure production	Automate as much as possible	SCC, PE, DEA, LOC
Paper submission	Meta data, file handling, resubmission	PE, POS, DBA, SCC
File management	Secure system for paper processing	SYS, DEA, PE, POS
Meta data management	Enter/edit/track	DEA, PE, POS
Paper processing management	Track/log all processing	POS, PE, DBA
CD/wrapper/WWW/hard copy	Extract all final papers and data for page numbering, adding conference identifier, create Web site, etc.	PE, DBA, SCC
Statistical analysis	Extract data, semi-automate	PE, SCC, SC, OC, DBA
Author interface	Turn off access once paper processing starts Suppress resubmission, unless requested Post final version on WWW	World
E-mailing	Global/interest group/specific e-mail lists from author profiles	PE, LOC, SCC, DBA
Itinerary planner	Personalized conference programme WWW and PDA support	DBA, PE

Editing a profile When the database is first launched, all authors will be invited to access their profile and then pose a question and provide an answer which only they should know. This is a similar system to that used on many e-commerce sites and would be used in place of a password or PIN number which authors would tend to forget if they only use the system once or twice a year.

In order to edit ones profile, the author would be presented with the posed question and only given access to the data if it is answered correctly. Other users would not be able to display any of the data.

The author database would contain the following fields, some of which would be uneditable as defined below.

Last name
First name or initial
Middle name or initial
Publication name preferred format to appear in the author list which would simplify the problem of accommodating initials, hyphens, double-character initials, Jr./Sr./I/II/III, etc.
e-mail address
Telephone number
FAX number
Affiliation only available from a pulldown list. The full institution name and shortened acronym (automatically generated from the database) to be used in various database and conference applications would be uneditable. Provision would have to be made to allow a new institution to be entered. But this would have to be vetted and approved by the database administrator, then added to the official pulldown list.
Street automatically filled based upon affiliation, uneditable.
City automatically filled based upon affiliation, uneditable.
State/Province/County automatically filled based upon affiliation, uneditable.
Post Code/ZIP automatically filled based upon affiliation, uneditable.
Country automatically filled based upon affiliation, uneditable.
Department author entered.
Mail Stop author entered.
PO Box author entered.
APS membership.
IEEE membership.

EPS membership.
Other membership – conference specific.
Temporary E-mail address plus begin and end dates in yymmdd format.
Temporary Telephone number plus begin and end dates.
Temporary FAX number plus begin and end dates.
Temporary Affiliation from pulldown list plus begin and end dates.

Unresolved problems include how to handle foreign accented characters which may be parsed using an extension of the PAC developed special font dictionary, or through the use of Unicode.

The initial creation and editing will be an onerous task.

A standard list of affiliations should be created from existing databases, using QSPIRES as a resource for addresses. The majority of the address fields should be automatically filled and uneditable in order to impose standard formats.

New entries and any profile editing should result in an e-mail being sent to the database administrator for checking and approval.

The issues of data privacy and granting permission to use e-mail addresses were raised. It was thought that a published declaration that the data would only be used for JACoW related purposes would suffice. If provision were made for indicating areas of interest, then perhaps customized e-mail distribution lists could be produced to disseminate information about related workshops, collaborations and activities.

For the most part the database would be self-updating as many authors consistently publish at JACoW conferences. However, this would only apply to the primary authors. Therefore a compulsory review and renewal may be necessary every three years.

Abstracts Breakdown of information and sources for the handling of the abstracts.

Author supplied information

Role: Submitter, Primary Author, Presenting Author, Co-author, Chairman.
Presentation Method: Invited, Oral, Poster.
Classification: Subject, Subtopic, SubSubtopic.
Title
Text with any references in the running text.
Acknowledgements Funding agency, contract # only.
WWW URL For author's home page or further information related to abstract.
Network Connection required for poster session.
Keywords 5 keywords from pulldown menu.

System generated information

Paper ID = Session ID + Sequence #.
TOC page #
Abstract ID#
Receipt Date/Timestamp.
Last Update Date/Timestamp.
Withdrawn Date/Timestamp.
Status Flags.

Programme editor supplied

Session: Chair, Place, Date, Time, Title, Type.

Papers Breakdown of information and sources for the handling of the papers.

Author supplied information

Platform
Software used
Deleted file Flag.
Files Full set: Document PostScript file, document source file, figure files.
Resubmissions Overwrite originals before processing.

System generated information

File #
Filename supplied
Filename used
Timestamp
E-mail paper status to author at conference. Alerts author if green dot goes back to red dot.

Programme editor and staff supplied

Processing Activity Log data.
Flags Many.
Status Green/red/yellow/brown.
Page #
Volume #
QA Flags Many.
Poster police flags Presented, manned, acceptable format – 3 yeses = okay.
Copyright Received.

Slides Author supplied and system generated information as for papers.

Oral session video Programme editor supplied file indexes.

Paper reception office All required functionality already covered under description of papers.

Other flags Provide the facility to add extra flags and fields as required.

OVERALL SCHEMA: PROCESSES AND DATA

Items originally intended to be covered in this session were essentially covered in the prolonged previous session.

Other issues

Pre-flight distill + Pitstop Text Box + Type 3 font check.

Abandon idea of doing this and e-mailing resulting .pdf file back to the author.

Investigate doing this (distill on the correct PC or Mac platform using watched folders) and making the resulting .pdf available to the processing office editors. If problems exist a log file will also be generated.

The paper submission process explicitly requests authors to submit a PostScript file of their paper. Try to develop the following procedure:

- check for correct file extension: .ps, .PS, .prn, .PRN;
- copy to file test.ps;
- auto distill on the correct platform;
- make available to proceedings office editors (see above);
- if no .pdf file produced, submitted file was not a PostScript file (may be pdf);
- send e-mail to author instructing them to submit a PostScript file;
- Note: still need document source file and figure files.

DESIGN AND IMPLEMENTATION SESSIONS

Database design

During the workshop the relational schemas and functionalities of the two databases used for EPAC'02 and PAC'01 were presented and discussed. Moreover, the changes that are taking place for the PAC'03 database were also added to the discussion.

As a general result, it appears that the two databases are very similar in design and global functionality. This is a positive discovery, so that a future merge is really feasible. Both EPAC and PAC databases can manage the data related to the authors of contributions, handle their contacts and roles (main author, co-author, etc.). The list of affiliations and countries are also managed, but it has been seen that letting the authors to freely enter this data creates a considerable overload on the editorial team, since the result is a heterogeneous and non-consistent list of affiliations (each possibly replicated with different addresses, or even just written down differently).

Obviously, the two databases handle the management of abstracts and contributions. In particular, they both provide procedures to create classifications and sessions: however, the terminology used (i.e., what exactly a classification and a session are) is different, and the adoption of a common standard is hence needed.

The databases allow the authors to upload their abstracts, view them via web, submit the contributions together with all the needed meta-data. Moreover, they both manage the entire Paper Processing course (paper checking by an Editor, author feedback through response dots – basically green or red – and paper correction/resubmission, re-editing, cropping, Quality Assurance, page numbering etc.).

There are, however, some differences or, better to say, features peculiar to each database.

The database that was used for EPAC'02, developed at CERN, implements a comprehensive logging of any action users perform via the web interface. In this way, it is possible to know when an author uploaded his or her files, when and who took them for processing, who entered any modification in the database data, and so on. In case a problem arises, this logging facility will furnish all the information needed to identify the problem (in quality and time) and to help for its solution.

To reduce any potential editor mistake in file down- and uploading during the paper processing stage, a separate application (developed in MS Visual Basic) integrates the database. An editor only needs to input the paper programme code: this utility will pick up the files from the file server, store them on the local machine and, when finished, help storing back the files to the server. Each time these files are uploaded, a new "version" is created, by storing them in a new "document directory" identified by a progressive version number: the old data, hence, will not be deleted.

What just stated indicates that the papers are stored on a fileserver, and only "referenced" by the database. This approach was initially thought as non-optimal from the developers of the PAC'01 database. One of their goals, then, was to store the papers directly in the Oracle database, together with the metadata and the general information for the conference. This leads to the possibility of bypassing the need for a separate file server (or service, if installed on the database server machine) for the conference: however, this loads the database (server) of more work. Due to bugs to that database software and to the fact that there were several authors trying to upload their papers in the very last hours before the electronic submission deadline, the system crashed, and the file upload was redirected to an emergency ftp server in a very short time.

By a post-conference analysis, it has been decided that an in-database paper storage will not carry to any new sensible improvement, and that a separate storage on a file server is preferred for future conferences.

During PAC'01 it was seen that communicating with the authors took a lot of time and work: for example, it is commonly needed to find out who did not submit all what requested (files, copyright forms, metadata etc.), so an editor has to build a list of those persons and then contact them via e-mail. Since this is a repetitive task, in the development version of the PAC'03 database a new web facility that helps e-mailing those persons in a semi-automatic way has been added.

Even if those two databases are very helpful for conference proceedings handling, they lack in a few aspects that have been pointed out during their use on the field. In particular, they do not cover delegates management, but focus only on papers and related authors. On this topic, however,

they don't provide support for authors with multiple affiliations and, in particular, for authors that present different contributions using different affiliations (e.g., a person that works on two different projects for two different laboratories, or one laboratory and a university).

Finally, even if they are of valuable support for paper processing and for the creation of the abstract booklet, CD-ROM and proceedings on paper (from now on, 'publications'), the process for this creation is based on the data extracted from the database and used by external procedures or programs (Visual Basic for Applications in Excel, or Visual Basic programs, or PERL scripts, or . . .).

New needs

During the paper analysis and processing by the Editorial Team, any editor has to get the files from the fileserver, do his/her work and then put on the fileserver the new ones. This process has usually been accomplished through a direct connection to a network share (via Microsoft Network, NFS or Appleshare) on a fileserver, or through a separate application as for EPAC'02. This process is really prone to errors because, apart from a certain level of logging, an editor could, for example, download the wrong files, or the wrong version of the right files, or upload the right files in a bad place, or forget to upload at all. Identifying such errors is a very complex task, with no success guaranteed. Hence, a new, completely automatic, way of getting these files from, or putting them to, the fileserver is requested. This could be done directly via the web interface of the database.

The whole stage of paper processing and creation of final publications needs to be developed, trying to automate it as much as possible, and letting the database to play a more important role in this phase.

Some conferences offer valuable registration fee reductions or particular conditions for APS or EPS members. When the database is expected to handle the registration of both authors and delegates, this feature will be needed. However, it has been decided that delegates management is at low priority at the present time.

In this direction, many current conferences (not in JACoW) handle participants' itinerary, i.e. furnish an automatic creation of session schedules together with places by only giving a list of chosen presentations. This really is an appreciated plus for a conference attendee. Then, the whole programme and this itinerary should be made available for download to the ever more spreading handheld devices (PDA's).

Finally, it could be useful to notify the authors about their paper processing status, via e-mail and during the conference also, in addition to the classic 'dots' way.

SOFTWARE DESIGN

PAC'01 DB

The PAC'01 Proceedings team based the conference information management on an Oracle 8 database on a machine running Microsoft Windows. The programming language used for the application was mainly Oracle PL/SQL stored in packages into the database; in addition, Oracle Forms were also used, but not with a web interface and only for convenience due to their presence in the installation. The web server was totally based on the custom Oracle Application Server, installed on the same machine.

EPAC'02 DB

Quite similarly, the EPAC'02 team adopted Oracle 8i, but under Solaris (on the standard Oracle server in CERN). The languages used were PL/SQL together with Active Server Pages (ASP) for the web application: the web server was Microsoft Internet Information Server (IIS) version 5.1 running on a separate Windows machine. On that server it was needed to install a free ActiveX component to handle file uploading.

PAC'03 DB

For PAC'03 it is planned to use Oracle 8i running on a Unix Operating System. The other characteristics will be much similar to those of PAC'01, being an evolution from that base. In particular, the language will still remain PL/SQL in database packages, but the web server will change to the new Oracle iAS, from the Oracle 9i suite, installed probably on the same server. This new Application Server is based on the Open Source Project Apache [1] together with the PL/SQL interface module (`mod_plsql`); a PERL CGI is under advanced development, to easily manage file uploading in place of ftp.

THE JACOW DB

These two parallel projects demonstrate the necessity and usefulness of having such a facility available for our congresses. Hence, is has been seen the opportunity to join the efforts and go on a common road with a unified database that hereafter will be called "the JACoW DB". This database will be made available for all the conferences in JACoW[1] and will consist in a central repository storing all the "people profiles" and all the data for the institutes. The place where to set up such a central repository is yet to be decided, since such a decision involves the approval of the possible hosting institution. This database could also contain all the data concerning the single conferences, although it would not be available on the web. Instead, every conference will have a copy of this standard JACoW DB

[1]It could also be made available for *any* conference, with no JACoW restriction. For this and in aim to also protect this software, I suggest to adopt the GNU General Public License (GNU GPL, http://www.gnu.org/licenses).

with, at the very beginning, a default set of reference data. During the whole conference process, this 'local' database will be populated with all the appropriate data and, after the process' conclusion, this data could be pushed onto the central JACoW DB. In particular, the conference paper metadata and all the software packages (PL/SQL) will pertain to the 'local' database, even if it will be inherited from the central JACoW DB.

Modules, requirements and assignments

Christine Petit-Jean-Genaz has volunteered to populate with data the centralized part of the JACoW DB, the Profiles and Institutes tables, with the information gathered from the past conferences. This is "good data", since it has already been normalized, with spurious duplicates merged.

Having a local (per-conference) database installation implies the need to have a database expert (DBA, or Data Base Administrator) available also locally. Finding such a person will be up to each conference's organization.

For the present time, the management of delegates in addition to authors is left in the wish-list, having other higher priorities.

Matt Arena will carry on the development of the module that will manage the submission of contributions via the web. Together with Pascal Le Roux, he will also develop the editor interface and the file management for the processing phase of the job.

Sara Webber will lead the development of the 'publications' module. Together with Pascal and Christine, she will take care of automating the Abstract Booklet and Proceedings creation stage, including the Table of Context, the Author Index, the page numbering and so on.

Inheriting from the PAC'03 DB, Matt will add the semi-automatic e-mailing feature to the JACoW DB.

The statistics and analysis part will be followed by John Poole: he will ask the PAC'03 team for plotting out what information would be found useful to get from the whole process and, hence, from the database.

Export Every new conference will have to install a local database from the central one, together with a copy of the repository. Exporting capabilities will be needed for this task and, in particular, two different 'exports' could be draft:

- small conferences could not have the appropriate resources to buy an Oracle license and to find a DBA to manage it. As an alternative, they could use other custom databases, using their particular local know-how. However, they will use the data stored in the central JACoW DB as a starting base. John will hence create the procedures to export this data in a standard and portable format;

- a full installation of both Oracle and the JACoW DB could be not a trivial task. Matt will provide a sort of "take-away package", or a CD with simple procedures to follow to do the whole job. Jeff Patton and Martin Comyn will help acting as install-testers to check for easiness and functionality.

Advanced Features Ivan Andrian will integrate Matt's work with a routine to check for correct uploaded file type, to ensure what we get from authors is true postscript.

Some things will be added during the general development, so no assignment to any person has been done. Among these are the creation of a personal itinerary, its download (together with the whole programme) do PDA's and the notification, via e-mail, to authors of their paper status.

The consideration of APS and EPS membership will be left in the wish-list by now, since it is related to the management of delegates (also left out for the moment).

MILESTONES

It has been decided to work towards a beta version of the JACoW DB by September 1, 2003. By this time, the database will be ready for final testing with tables populated. After a testing slot of a month, on October 1, 2003 the JACoW DB will get to version 1.0, ready for its use on-the-field.

To better monitor the development process and to coordinate all the modules/developers, a technical review meeting has been fixed for just before PAC'03.

John will coordinate all the development process, collecting the documentation and letting it circulate among the other actors. To better communicate, a mailing list will be set up shortly.

JACOW AND INDICO

During the workshop it has been discovered that both JACoW (DB) and InDiCo [2] are two "new projects" that have many things in common. However, there are some peculiar aspects for both: JACoW has gained a lot of experience on the field, and now knows very well all the real problems related to the conferences in it. JACoW, also, focuses very much on the paper processing, where InDiCo is not interested (at least, in the way JACoW does). The community that JACoW relates to has become very large, and is increasing as new conferences are joining.

InDiCo, on the counterpart, has a lot of scientific experience on the background, starting from where other previous projects finished and hence inheriting a lot of knowledge on, for example, advanced tools and techniques for Information Retrieval, on multimedia inclusion and management and so on [3] [4] [5].

It is still an open question if the two projects could ever be integrated one in the other. To try answering such a question, it has been decided to keep in contact, to exchange experiences and, on the JACoW side, to contribute

to InDiCo with requirements and possible issues, in an effort to help its development with ‘real’ situations. It will be important to study its modulability, that’s one of its fundamentals, so to try integrating it with a “JACoW paper processing module”, if possible.

In this direction, Ivan will help keeping the contacts among JACoW and the Italian developers of InDiCo, at SISSA [6], as the JACoW people from CERN will do for the CERN counterpart of this project.

REFERENCES

[1] http://www.oracle.com/ip/deploy/ias
http://www.apache.org

[2] http://www.cern.ch/indico

[3] http://tips.sissa.it

[4] http://torii.sissa.it

[5] http://jhep.sissa.it

[6] SISSA / ISAS – Scuola Internazionale Superiore di Studi Avanzati / International School for Advanced Studies
http://www.sissa.it