The Newsletter for SIR UK Users

SIR REPORTER

Contents

Editorial

As many of you know, Kathy Brooks usually writes the Reporter editorial but as ever at this time of year she is buried in school league tables and is unlikely to surface much before Christmas. Let's hope the schools' tables don't cause the same sort of chaos as the 'A' Level results have done!

As the evenings draw in rapidly, this issue of Reporter looks back to those balmy summer evenings in the form of a report from the SIR Users' Conference held in Aberdeen at the end of June. It was good to see so many people there – both old and new faces. Aberdeen may not be able to boast temperatures to match the Mediterranean, but at that time of year it never quite gets dark and it was really nice being able to walk back to my hotel at 2:00 a.m. in twilight.

Before the conference there were two days of SIR training held in two parallel sessions – introductory training for people new to SIR and advanced training for the old hands. I'm sure that everyone who was there would agree with me that both sessions were very successful and will join me in thanking Tony Reardon, John Lemon, Tom Shriver and Dave Doulton for their contributions. When you think how much Microsoft charge for their courses, you realise what good value SIR conference training is!

Many thanks go to John Lemon (again!) for organising the three days, taking us on a guided tour of his home town and, despite being a non-imbiber himself, introducing us to the delights of some rather exclusive Highland single malts.

Fran Williams
Fwill@essex.ac.uk

SIR is at http://www.sir.com.au
UK SIR Users Group is at http://www.soton.ac.uk/~sug

Note from the ChairFollowing the AGM at yet another successful conference, I have again been elected Chair by the committee so you will have to put up with my comments for another year at least.

I would like to give thanks here to two members of the committee who did not stand for re-election due to the changing natures of their work and one whose circumstances changed due to the birth of a child. They are Charlie Owen, who has been a member of the committee since election at the 1993 AGM, and Michael Staley, who beats him by a year being elected at the 1992 AGM. Michael's regular contributions to SIR Reporter will be missed by us all. We may be able to twist his arm to produce something as a non committee member. Lisa Zanninetti was a relatively new recruit to the ranks, only being elected at the AGM of 1997. Between them we are losing at least 25 years of SIR and committee experience which will be hard to replace.

At the AGM there were no nominations for new committee members, so we are short a few committee members. If any of you readers feel you can contribute to helping us organise successful conferences and interesting SIR Reporters then please let me know. Even though the next AGM is not until next year, the committee is allowed to co-opt people to serve; so please act soon - do not wait until next year.

Dave Doulton

SIR Users Conference June 21st 2002

Atholl Hotel, Aberdeen, Scotland

Breakfast in the Last Chance Saloon! - by your special correspondent Golden Balls (Adrian Hodgson)

The world cup shoot-out featuring Big Phil versus Smooth Sven.

I could say 'we was robbed' but we weren't really.

England's usual line-up struggled to get sufficient possession of the ball but briefly raised all our hopes on 23 minutes when an exquisite through ball from Heskey caught Lucio out in the Brazilian defence, and our man Michael (complete with groin strain) planted the ball firmly in the net.

Now all we had to do was keep another clean sheet.

Unfortunately Rivaldo (47mins) popped up on the right wing just before half time and snapped up a low driven equaliser into the bottom left hand corner. Then just after the half time pep talk from Smooth Sven, Ronaldinho (50 mins) took a bizarre free kick that turned into a fluke shot from 40 yards out.

Now all we had to do was score two more goals! After 58 minutes we had the opportunity of doing this against ten men, as Ronaldinho was given a straight red for a nasty follow through on Mills. Hopes were briefly raised again. However, most of the teams creativity and get up and go seemed to be 'lost and gone' despite bringing on the fresh legs of Dyer, Vasell, and Sheringham, and we proceeded with the usual collection of lame crosses, and getting caught in possession.

The final nail in the coffin was the broad beam on the faces of the Scottish Atholl staff and the frantic tooting of cars in the street outside. Were we really in Aberdeen or had we somehow found ourselves in Rio de Janeiro?

Keynote address: SIR 2002

Tony Reardon, SIR Pty Ltd

The overall theme was: -

- Where have we come from?
- Where are we now?
- Where are we going?

Tony started by talking about the early days.

Version 1 on the Control Data, with Kathy the only user present at the conference.

Version 2 that started life as a multiple - platform product from Day 1, with separate modules for SQL and Forms and introduced the famous X command.

At this point there was a brief 'chronological' interruption from the 21st century as one of the modern laptops announced in a female voice

'Batteries are fully charged' accompanied by much hilarity!

Version 3 then arrived with the emphasis on PQL and building applications for other users. This was the era of text screens, menus and function keys with Forms and SQL continuing to be developed as separate modules.

Tony then brought us up to date with the step change to SIR 2000/2, and the facility to have multiple or indeed no databases attached. This was the first time in 20 years that SIR could be used without a database attached, and has seen the introduction of secondary indices which Tony described as 25 years overdue. Other recent developments include integrated Master modules, SQL server for ODBC, the long string facility (4094 characters), built in memory management, the Graphical User Interface approach of visual PQL including screen painters and PQL forms, spreadsheet views of the data, and internet capabilities.

The product now has a single source code under development on Windows XP with approximately 470,000 lines of code and over 2300 modules written in C, C++ and PQL. Releases are available for Windows, Sun, HP, IBM, Linux and Open VMS, with only small differences between these releases. The strategy for fixing bugs & features has changed with most issues fixed straight away rather than waiting for a major release.

Where next?

The strategic aim is to develop a fully integrated product that incorporates PQL Forms and SQL within the core of the software. It is intended that visual PQL becomes the primary tool and that the old text style screens, old Forms, old editor, PQL debugger and tabfiles are removed

Next interim release

This is currently scheduled for Spring 2003. It is likely to include a computer based training module for self-learning, extended GUI and paint features to help support HTML format documentation and extended support for alternative image file formats

Next major release (2004?)

- Incorporate SQL in visual PQL
- Extend and improve backup and restore
- Multi-database data files (i.e. facility to split the .SR3 file into chunks)
- PQL server
- XML input /output
- A GUI de-bugger (possible)
- 32 character variable names (possible – but a major piece of work)

SIR as the Foundation for a Health Sciences Data Coordinating Center

Howard Andrews & John Pittman, Columbia University, USA

Howard commenced his talk with a summary of the history of the use of SIR at Columbia University

1981 – John Pitman purchases SIR V2. and starts to use for research projects on IBM.

1988 – The team is known as the Statistical Analysis Unit (SAU) with 4-5 staff. They conduct their first external project using SIR; a longitudinal study of Alzheimer's disease for 2000 residents of northern Manhattan.

The team succeeds in having a grant renewed for five years

1995-1999. Renamed as Data Co-ordinating Centre (DCC) with 6-8 staff.

Asked to support work for the HIV centre and the Centre for Child Environmental Health.

2002. The team now supports 25 databases using dual processor hardware on a 512 MB RAM server, with three 18GB RAID drives under Windows NT

The success factors were: -

- Experienced data management staff
- SIR maximises power and minimises cost
- Other Universities try an uncoordinated non-central approach using inexperienced graduates with SAS and SPSS. Each department does its own thing.
- Centralisation of data management creates a favourable climate for SIR
- Accountability a key driver with each grant for new work supporting DCC staff
- SIR delivers data in the format that researchers need (e.g. SPSS)
- Integrated variable and value labelling. User defined missing values can be ported to SPSS and SAS in self documenting, analysis ready statistical files
- Planned approach to data entry with input to form design and procedure for notifying form changes to DCC

New developments

The future looks bright for SIR in medical research, as there is a large untapped market to provide centralised support at Universities. Researchers are frequently looking for advice about data protection and confidentiality, which again can best be supported by a centre of excellence.

The DCC is aiming to run courses in data management starting in autumn 2002.
There is a growing need to make data sets more available via the Web with a better standard of labels & documentation

XML & SVG from PQL

Dave Doulton, University of Southampton, UK

What is XML?

eXtensible Markup Language – a superset of HTML. Essentially a way of marking up data to impose a hierarchical structure using mark-up language tags, that is useful for web applications.

e.g. for the Company database

```
<COMPANY>
   <CASE>
      <EMPLOYEE>
        <name>
           John D Jones
         </name>
```

What use is it?

Avoids the pitfalls of comma delimited files (e.g. embedded commas), and it's the way forward.

Snags - <> & ' and " need replacing in data first (e.g. &gt). Also in PQL the tags <> look like global variables.

In order to make use of data in XML format, need to be able to read and manipulate the data using either

- XSL (eXtensible Stylesheet Language)
- SVG (Scalable Vector Graphics) – to do this need to download an SVG plug in from Adobe.

XSL has style sheets that enable the XML data set to be formatted in alternative styles. Opening an XML file in a Web Browser is helpful because the Browser provides an indented layout to highlight the hierarchy.

SVG is used for XML files that represent a series of drawing instructions (e.g. a histogram). The instructions can include details of different animation styles for drawing parts of the picture. (i.e. draw a bar in red from point A to Point B taking 30 seconds to do it)

Example application:

The Administration at Southampton University wants to construct an emailing list from Dave's SIR database. They send a WEB request to the SIR database, which returns an XML data screen, which can then be used to build the e-mailing list

Asbestos Proficiency Testing at the Institute of Occupational Medicine

Patrick Brown, Institute of Occupational Medicine, UK

The I.O.M undertakes research into Occupational and Environmental Health Issues. Its roots lie in work undertaken between the 1960's and 1980's to research coal miners' illnesses. Research is now being undertaken to test the accuracy of critical laboratory measurements of airborne asbestos fibre concentrations from the 200 labs in the UK

The approach involves the establishment of reference values for slide samples, and then the analysis of each sample by several laboratories. Those that do well in relation to the reference values are published on a list. The testing takes place quarterly and 2,500 measurements are used each time.

A SIR database is used with record types for labs, samples and rounds (i.e. quarterly testing)
The database has 300 cases, 20 record types and 200,000 records. Sir/Forms and Excel worksheets are used for input.

Conversion from Sir 3.2 to Sir 2000.

This went smoothly and was helped by the prompt help from SIR Pty whenever any issues arose. SIR Pty were particularly helpful with the conversion of Patrick's menu system and the use of commands such as EXECUTE DBMS 'CALL DATENTRY.RES1:T'

The worst problem encountered was a bug in the Unload/reload routine for SIR2000, which resulted in the loss of CIR variables on occasions.

Standard dialogues have been used to replace screen input.

One issue that did arise was the temptation to continue ad-infinitum adding validation to data input dialogues. In retrospect it would have been better to plan this work first.

An unexpected side benefit of the transfer was the opportunity to clear out junk or redundant

programs from the procedure file. It was also encouraging that the end users did not need much additional training after the conversion.

Potential future developments include the use of ODBC, PQL forms (SIR 2002), electronic reports, and the creation of a web site.

Lookups; the Good, the Bad and the Ugly.

John Lemon, University of Aberdeen, UK

As ever the data set is the Aberdeen Maternity and Neo-natal databank which has 34-40 record types for 120,000 women.

Lookups are needed for coded text. E.g. the label 'Uterine adhesions' which requires 19 bytes to store as a label has an associated code of 621.5 which only requires 8 bytes.

Lookups in the early days were mainly needed for storage space efficiency. These days the main reason for using lookups rather than value labels is for lengthy lists that need to be maintained and updated.

SIR version 2 and the dummy case.

In the era of SIR version 2 the only option was to use a dummy case in the database e.g. Case 0 or Case –999999.

The problems with this were: -

- Only available for one database, repeat for all databases needing lookup
- Needed inverted lists to lookup reference in the opposite order
- If chose a text field as a key added extra bytes per record due to the effect of the MAX KEY SIZE setting.

SQL tab files

The next approach was the use of SQL tab files.

Advantages

- One external file
- Multi-user file for all lookups
- Simplified maintenance
- Removed issue of MAX KEY SIZE

Disadvantages

- No equivalent of database file dump /add records
- Tab file can easily be corrupted
- Export needed to be used for backup

SIR 2000 and the second database

One of the advances with SIR 2000 was the facility to connect more than one database. However this was still limited for lookup purposes as there was no alternative views or indices available. John didn't use these features because of this.

SIR 2002 and secondary indices

The latest innovation is the combination of multiple databases with secondary indices and the Lookup command. John has set up a second caseless database with lots of secondary indices. The key fields are numeric codes to keep the MAX KEY SIZE for the database small.

John described progress on this conversion as slow as he continued to encounter 'another SIR feature'! At times he felt like a snail in a well progressing 3 feet up during the day, only to slip 2 feet back at night. As the Goons would say, 'he's fallen in the water'. Eventually, with plenty of help from SIR he has cracked it, and come up with a final solution that looks like the way forward for lookups.

Progress testing with SIR

Dr David Keane, McMaster University, Canada

What is it?

Testing student's knowledge acquisition and retention over time in relation to curriculum objectives. I.e. Are you learning enough, retaining enough and doing it quickly enough.

The purposes of this are: -

- To help the students by providing feedback in relation to self-assessment.
- To collect evidence to support marginal pass /fail decisions.

Method

An item bank is created with samples of the knowledge that a student will have encountered both by the time they graduate and six months beyond graduation.

To enable a true measure, it is important that the students don't study for this test and don't guess the answers.

A SIR database has been created to store the questions and identifiers and the correct responses. It holds a large volume of text information, including special characters such as Greek letters and Maths symbols. The students' results are scanned and available as a tab

delimited ASCII file. Results breakdowns are required by class, examinee etc.

One case structured database ITEMS holds the questions and correct responses. A second case structured database called TEEX holds the test results, examinee and class details.

To help with the text intensive requirements, Word Perfect report templates were created with data merged from csv files downloaded from SIR

What David values about SIR?

- Case based database and utilities
- Global variables
- File Dump /File input
- Tabfiles and tables
- PQL – nested access to cases
- Visual dialog boxes
- Easy generation of SPSS system files

Future developments

These revolve around SIR 2002 and the use of secondary indices, Web Access and CGI scripts

National Diet and Nutrition Survey

Michael Staley, ONS, UK

This survey involved a questionnaire in Blaise together with what Mike referred to as 'bolt – on goodies'. These add-ons included blood tests, urine samples, blood pressure measurement, an activity and dietary diary and a bowel movements diary!

Consistency checks were built in. For example from the questionnaire if the respondent said they didn't drink tea, and then the dietary dairy showed 20-30 cups per day, this would not only be inconsistent, but we would expect a good urine sample as well!

Data was edited in Clipper and merged into the SIR database. The Blaise data was passed through a facility called Manipula to format the data ready for SIR.

Derived variables such as nutritional measures, and age of respondent at certain points were created in SIR. The final data sets were sent to the end users as SIR export files.

A baker's dozen in a button

Tom Shriver, DataVisor, USA

Tom described how we all write the same or similar lines of PQL code over and over again. The temptation is to do this by the cut and paste method. Tom has developed a more generic approach where he has created a template with standard data prompts, such as filenames.

The process of developing this baker's dozen tool went as follows: -

- Generate a PQL template
- Edit it and make it the default file.
- Enables you to run many programs at once.
- Provide user prompts for data ranges (record types, case ids etc)
- Add facility to abort complex procedures mid-stream if they crash
- Disconnect and re-connect MASTER for speed
- Use templates for PQL output
- ? (sorry missed this one – Adrian !)
- Edit output
- Print the output
- Display as a Web page
- Output to SPSS
- Put all of these facilities into a button

Tom also added the facility to enter date short-hands (e.g. 1 1 2 for 1st January 2002)

The post processing of the PQL (e.g. printing output files) is enabled only if the PQL runs correctly. Tom used a system Global called TERROR (Task Error) to check for this.

Using SIR for survey research

Klaus Allerbeck, Johann Wolfgang Goethe Universitae, Germany

Klaus described how they had used SIR for several studies. The first was a study of German youth in 1983, with re-interviews of the same respondents in 1997. Another study was a sample of Boston Computer Society members, conducted from Frankfurt and completely handled by SIR/DBMS and SIR/FORMS over ten years ago. Of course, the ability of SIR to produce system files for different analysis systems was used, as well as its ability to produce camera-ready tables.

Survey data are special in three ways: for many variables (such as party or brand preference etc), there are numeric codes without intrinsic meaning - thus, value labels are needed. The data may be missing for various reasons - the question may be inappropriate, the respondent may not know, or might refuse to answer - thus, missing values and appropriate procedures to handle them are needed. The data are "almost" rectangular, as in a study where there is previous information about both respondents and drop-outs or in panel studies. The "respondent" or "prospective respondent" is a rather natural

CASE. These three aspects make SIR a rather suitable data management system for survey researchers.

In conclusion SIR seems to be the system of choices for surveys today. For older data sets that may exist in punched card form only, some extensions seem to be appropriate to reduce dependency on poorly documented legacy systems.

Tips and Tricks

A handy tip is to explore sysproc members to see how things are done in the user interface, if you see something there you want to use have a look at how it is done and call the code to do it for you. For example I wanted a directory browser for a little program I was writing and knew there was one in the open database dialog. I went looking for it and found SUBROUTINE SYSPROC.MENU.DBROWSE which take a parameter of where you want to start in the directory/folder tree and returns the directory selected.

I then used it in my program, which some of you might find useful. I have a program that when you scan in images calls them Untitled1.jpg Untitled2.jpg etc. however next time you use it it starts at 1 again which overwrites the originals. I wanted a quick way to rename them after scanning to avoid the problem I rename them pic01.jpg pic02.jpg etc assuming there are not more than 99 in a folder they will now sort in the right order by names. I wrote the following pql which asks for the directory and the number to start renumbering from and then renames the files. I had thought of searching for the next pic number to start at in the program, but found the need to rescan an image and wanted to fill in an existing pic number so gave the option of supplying the value as a default.

The program is as follows

```
program
string*1000 path,mask,fn,nfn,starts
EXECUTE SUBROUTINE SYSPROC.MENU.DBROWSE
   ("c:\") RETURNING (PATH)
if(len(trim(path))=0 or trim(path)='\')exit
   program
compute mask=trim(path)+"pic*.jpg"
compute files=filecnt(mask)
compute m=0

for i=1,files
. compute fnum=numbr(sbst(filen(mask,i),4,2))
. if(fnum > m) compute m=fnum
end for

compute m=m+1
display textbox 'start at number (default
   '+format(m)+')' response ok,starts
ifthen(ok=-1)
. compute start=m
else
. compute start=numbr(starts)
. ifthen(exists(start)=0 or start eq 0)
.   compute start=m
. endif
endif

compute mask=trim(path)+'Un*.jpg'
compute files=filecnt(mask)

for i=1,files
. compute fn=filen(mask,i)
. compute j=abs(srst(fn,'.'))
. ifthen(j=11)
.   compute fnum=numbr(sbst(fn,j-1,1))
. else
.   compute fnum=numbr(sbst(fn,j-2,2))
. endif
. compute fnum=fnum+start-1
. compute
   nfn='pic'+fill(format(fnum,2),'0')+'.jpg'
. compute fn=path+fn
. compute nfn=path+nfn
. write fn
. write nfn
. compute dummy=rnmfile(fn,nfn)
end for

end program
```

Dave Doulton

SIR UK User Group Committee Members 2002/2003

Randy Banks (Treasurer)
ISER
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
Tel: 01206 873067
Fax: 01206 873151
email: randy@essex.ac.uk

Kathy Brooks (Reporter)
Forvus
53 Clapham Common
South Side
London SW4 9BX
Tel: 020 7819 1012
Fax: 020 7819 1010
email: kathy@forvus.co.uk

Patrick Brown
Institute of Occupational Medicine
8 Roxburgh Place
Edinburgh EH8 9SU
Tel: 0131 667 5131
Fax: 0131 667 0136
email: Patrick.Brown@iomhq.org.uk

Dave Doulton (Chair)
University of Southampton
Computing Services
Highfield
Southampton SO17 1BJ
Tel: 023 8059 3541
Fax: 023 8059 3131
email: D.C.Doulton@soton.ac.uk

Adrian Hodgson
ORC International
5th Floor City Point
701 Chester Road
Stretford
Manchester M32 0RW
Tel: 0161 877 6781 (switchboard)
Fax: 0161 872 3997
Email: adrian.hodgson@orc.co.uk

John S. Lemon
Aberdeen University Computing Centre
Edward Wright Building
Dunbar Street
Aberdeen AB24 3QY
Tel: 01224 273350
Fax: 01224 273372
Email: j.s.lemon@abdn.ac.uk

Mo Reardon/Tony Reardon
SIR PTY LTD
312 Mona Vale Road
Terrey Hills
NSW 2084, Australia
Tel: 00612 9450 2354
Fax: 00612 9475 1430
email: mo@sir.com.au
email: tony@sir.com.au

Frances Williams (Secretary)
ISER
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
Tel: 01206 873568
Fax: 01206 873151
email: fwill@essex.ac.uk

You can email the whole committee by addressing your message to **sug.comm@essex.ac.uk**