Quan-products - communicating with the outside world

Paper given by Mark Katz, Mark-IT to the SPSS/MR user group meeting
September 5-6th 2002, London
V1.3b

Overview

Most of the papers at this conference deal with the new DIMENSIONS concept, which allows for swift development of tools and transfer of data/data-specification across Quan-products. There is, however, a significant market place out there of companies (agencies and end-clients) who do not use Quan-products and need to be able to pass data/specs both into and out of a Quan-products environment.

Sparked off by some pioneering work in the early 90's by a team of UK programmers and researchers, many international "standards" have now been established for handling survey data. This means that data/specs can be converted from any system to a standard format and, similarly, data/specs in this standard can be converted to the package of your choice.

One of the leaders in this MR field is **triple-s** http://www.triple-s.org/. Our company produced the **triple-s** to quantum converter in June last year and ATP (ironically - independently and unknown to us) produced the quantum to **triple-s** converter at the same time.

In this paper we expand the concept of standards and converters to look not only at **triple-s** but also a) The strategic concepts of using XML to provide a platform for embedding all aspects of survey information within an XML-based data/knowledge base and b) An allied, and integral, part of these converters – namely free/open software

Contents

- triple-s - what it is it how does it work
- The triple-s to Quantum converters
- Challenges of converting from Quantum to/from triple-s
- Opensurvey and tabxml
- Other initiatives for establishing standards

1. triple-s - what is it and how does it look

1.1 What is it?

triple-s is a language that describes survey data; this language is defined very mathematically using a concept known as XML. XML is closely related to the language used to drive the web and can be seen in any web/HTML page using a set of definitions called “tags”

Survey data that has been converted to **triple-s** comprises two parts

a. **Data** - a data-file holding all the data where each record represents the answers given by a respondent. The file is in fixed width format so that each question (or variable) may be found at a fixed location on the record

b. **Definitions** - this is the map that shows, for each question/variable where it is in the data file, its' name, question text, type of variable and any codes

For a more detailed technical explanation of the evolution of **triple-s** and its' definition, read "A standard within a standard" by Keith Hughes, Stephen Jenkins and Geoff Wright, ASC March 1999 www.triple-s.org/sssasc99.htm.

1.2 An example of triple-s

Let's give an example using some Quantum and triple-s

We have a variable SEX; with a title "What is your gender" coded into column 6 with a value 1 for Male, 2 for female and 3 for "not sure"

Under Quantum we would write this as

```
L sex
ttlWhat is your gender
n10Base
n01Male;c6='1'
n01Female;c6='2'
n01Not sure;c6='3'
```

[we could use "col 6 Base;Male;Female;Not Sure" but we need to show it more explicitly]

Under triple-s definitions this would be

```
<variable ident="2" type="single">
<name>sex</name>
<label>What is your gender</label>
<position start="6" finish="6"/>
     <values>
          <value code="1">Male</value>
          <value code="2">Female</value>
          <value code="3">Not sure</value>
     </values>
</variable>
```

Showing that this is the second variable and is single-coded. The name is sex and label is "What is your gender". The data starts and ends in column 6 and against each possible value (of 1, 2 & 3) there is a text of Male, Female, Not sure

1.3 Some background information

triple-s is a totally UK-based development – started in 1998, by a team of people under the aegis of the ASC (Association for Survey Computing) http://www.asc.org.uk/. ASC is holding a conference on Open Standards at their conference in 2 weeks here in London. A visit to their web site shows that some 24 converters (i.e. virtually all of the mainstream Survey processing systems) provide both input and output to

triple-s.

In summary it can be seen that if there are N different packages and we wish to achieve a bi-directional transfer between each of the languages then we need only 2*N converters instead of N**2 (N-squared) combinations. With almost 20 survey processing packages this is a tremendous saving.

The converters are freely available at the triple-s web site – including our triple-s to Quantum converter in various flavours. Our company has also written an Excel to triple-s and is working on a triple-s to SAS converters.

Work is almost complete on the Quantum converter for the recently announced V1.2 extension of triple-s, which handles filters, multi-language and weights etc

2. The triple-s to Quantum converters

The triple-s to Quantum converter (sssqt) is available at no charge as an EXE file for DOS/Windows. It is written in posix **awk** and the source code is also available allowing it to be run on a UNIX platform (or any other Operating system that supports the awk language).
It operates in batch mode – i.e. it takes the triple-s definition and produces a set of Quantum files (run, edit, tabs, axes etc - all linked with include files). The quantum command file would be

`quantum run data-file` (where data-file is the triple-s data-files)

It creates axes using ***n01***'s for conventional text variables and ***val*** for numeric ones – to include ranges. It also generates named ***inc***'s and ***alpha***'s for each of the variables. For multiple repeating fields it will generate a ***fld*** axis. It also creates LIST statements - if you wish to check the distribution of numeric and alpha field
There is an option to insert special statement for Quanvert/flip runs (e.g. call alpha).

A set of tab statements is produced for each of the variables of the form

`tab axis-name bkdown`

and it's up to the user to insert a breakdown of their choice.
It operates very quickly taking just a few seconds to convert the data-definition files. The triple-s data file remains untouched

3. Challenges of converting from Quantum to/from triple-s

3.1 triple-s to quantum

This is fairly straightforward, and the regular axes present no problems. Quantum cannot handle numeric and alpha variables without some manual intervention. triple-s does not allow (easily) multi-code variables that use just one column

All variable definitions of numeric fields are specifically put at the start of the **run** file, not a separate 'variable(s)', file in order to preserve UNIX/Dos Quantum compatibility

Two situation are highlighted if they occur

- Quantum is unable to handle the width of the **fld** field if it's more than 8 columns
- Not all title can be more than 200 chars or have continuation lines. These are truncated
- Quantum (DOS) silently fails to handle variables with names like **lpt1, con, prn,**

cd1 etc

Before attempting to compile the Quantum, the user must supply
- A file called BASE with an N10, N03Base etc
- A file called "bkdown" which is used for tabs file

3.2 Assumptions

The triple-s to Quantum converter makes some assumptions, i.e. it does not check the input XML code and may well fail if errors exist. It will give some indication as to which statement it failed on. It assumes

- All value fields are correct i.e. the length of a variable is precisely the difference between the *start* and *end* column
- All XML statements are correct with opening and closing tags
- All XML statements fit on just one line
- It tolerates blanks for text of numeric
- It replaces equal-signs ("=") in the row-text with 'equal'
- The html Entity References - &gt, &lt, &pos, &quot, &amp are translated. Other decimal codes are not.

3.3. VARIABLE/AXES NAMES

Old versions of Quantum limit the length of variable names to less than 8 characters. For this reason, the converter truncates any that are longer than 7. If there are any duplicates, it generates a temporary name and puts the details in a comment statement. Quantity/numeric items are treated in the same way

For the advanced version of sssqt, the following is included – they are designed to make it easier for Quanvert

- Includes a **rej=-** at the end of all axes
- **nz**
- means on **incs**
- Integer's are different to real's
- All **incs** are included on a **tab** statements
- **rqd** statements to identify/display erroneous data in the **EDIT** section
- a **CALL ALPHA** for each character field

3.4 quantum to triple-s

This works in the same way as **[n]qtspss** – in fact some people prefer to use this converter than the Quan-products program since it corrects a major error. This program will also create CSV and SPSS files..

The following situations should be borne in mind

- Quantum allows the user to recode variables in the **edit** section, so the definitions of the data are derived from the axes specification. This means that the variables may not be the same as the raw data (i.e it may have been recoded in EDIT section or codes joined together in the row definitions).

  The resulting triple-s data files have values that depend on the axis - 1 for the first row, 2 for the next – and a **don't know** is unlikely to have the '9' or '99' value.

- A new data file has to be created for triple-s – one that may bear little relationship to the original quantum input data-file

- Since it only uses the axes, any field on the data that is not included in the Quantum script, will not be included in the triple-s file

- Some variables/questions may be included twice, if they appear in more than one axis

- Within variables some individual values/text may be omitted if they are 'collapsed' in the axis – i.e. a text is allocated to 2 or more code with an *or* condition

4. Opensurvey and other standards

A common thread across all these converters is that the software is available and distributed free of charge! Many of the converters are distributed as part of the Open Source Initiative. Their model for the evolution of software is that the software is free but the consultancy; training and bespoke development is chargeable.

4.1 Open Source Initiative (OSI) is a non-profit corporation dedicated to managing and promoting the "Open Source " for the good of the community, specifically through the OSI Certified Open Source Software for more information about http://www.opensource.org/index.php

The basic idea behind open source is very simple: When programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, people adapt it and people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing. The open source community has learned that this rapid evolutionary process produces better software than the traditional closed model, in which only a very few programmers can see the source and everybody else must blindly use an opaque block of bits.

The concept of open source was brought to the UNIX world back in 1984 through the GNU initiative http://www.gnu.org/. They have a mass of software tools and enthusiasts who provide their time and the software at no charge

Perhaps the most prominent software products available under this scheme are

- Linux – the UNIX look alike
- OpenOffice http://www.openoffice.org/ and star Office http://wwws.sun.com/software/product_family/staroffice.html -designed to make all Microsoft office products available on a non-windows platform
- StarOffice software runs on multiple operating systems, including Solaris[tm] Operating Environment, Microsoft Windows, and Linux. The office suite has a simple, easy-to-use interface and contains full-featured applications including word processing, spreadsheet, presentation, graphics and database capabilities. Fully compatible with other office suites including Microsoft Office

In the last month, in an effort to achieve more efficient and cost-effective computing, the public sector has adopted Opensource "The use of the Open Source Software policy within UK Government should encourage the procurement of value-for-money solutions and lessen the reliance on individual IT suppliers."

4.2 Within our Market Research industry this concept of open source has been adopted by Opensurvey (http://www.opensurvey.org/) which aims to provide - at no charge - a platform and tools for transferring survey data/specs in an 'open' way between software packages. Apart from the design and documentation, they also encourage the free transfer of such information and tools.

This organisation - again lead and inspired by 'brits' - has focused on two standards

4.1 tabsml

TabsML addresses the need for a common standard format for crosstab reports. It defines very precisely the format of a table to identify such things as the page title, column titles, sub-titles, row text, overflow rows etc.

In this way, reports derived from different analysis tools can be converted for display and aggregation by different applications.

TabsML was pioneered by E-Tabs (www.e-tabs.com) in conjunction with OpenSurvey. For more information see www.opensurvey.org/ostabsml.htm

4.2 askml

The AskML project aims to produce a system independent XML based language for describing the content of surveys, particularly surveys delivered through Computer Assisted Interviewing. It focuses on the questions/answers and resulting activity and goes further than triple-s in three areas

- **Routing and conditional** - use of ***goto*** and ***if/else*** constructs
- **Expression Syntax** – to use 'arithmetic and calculate new values
- **Multiple use of blocks** - a collection of question prototypes

More information from http://www.opensurvey.org/osaskml.htm

5. Other initiatives for establishing standards for data-transfer

There are many international initiatives to standardise on metadata for the survey field with perhaps 3 leaders - the European Union is involved in two of them

5.1 Metanet
MetaNet - http://www.epros.ed.ac.uk/metanet/ - "a network of excellence for harmonising and synthesising the development of statistical metadata - is a part of the European Union Fifth framework Research and Development program."

5.2 Faster – Flexible Access to Statistics, Tables and Electronic Resources is a similar project with perhaps wider aims. http://www.faster-data.org/

5.3 DDI This project is based at the University of Michigan (http://www.icpsr.umich.edu/DDI/). The Data Documentation Initiative (DDI) is an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of "metadata" about datasets in the social and behavioural sciences.

"With the achievements of the DDI, codebooks can now be created in a uniform,

highly structured format that is easily and precisely searchable on the Web, that lends itself well to simultaneous use of multiple datasets, and that will significantly improve the content and usability of metadata. Further, this specification may have far-reaching implications for improvement of the entire process of data collection, data dissemination, and data analysis"

5.4 Nesstar

Nesstar is one of the first major initiatives to consolidate and display all aspects of a survey into a single 'package' built around the concepts of meta-data. It is being developed at the University of Essex (Data-archive) and the Norwegian Social Science Data Services (NSD) with funding from the EU. It brings together 4 issues and links them with hyperlinks.

Tools: Finding and sorting, Browsing, Analysing, Publishing
Text: Journal articles, User guides, Methodology instructions
Data: Micro, Aggregate, Time Series, Geographical, Qualitative
People: E-mail, Discussion-lists, Conferences, Expert networks

The overall design is build around the DDI model and has produced today a working model of all the above components. Current surveys include the high profile Labour Force Survey, General Household survey, Welsh referendum survey and Euro-barometer. Although currently slanted towards Social surveys and Policy makers (local, national and international), Researchers (academic, public sector and commercial) and General public (casual enquirers), I believe it has tremendous implications for the longer-term evolution of survey deliverables within our industry. For more information I well recommend a visit to http://www.nesstar.org/papers/ and in particular http://www.nesstar.org/papers/GlobalAccess.html
It is to be regretted that most of this research is in the academic/public sector.

Summary

These standards will bring a closer connection between the work we do in the MR world and the much bigger world of data-warehouses, Olap and cube technology – see http://www.omg.org/cwm/ This topic was covered, by the author at an ASC conference http://www.asc.org.uk/Events/Sep01/Abstract/Katz.htm or the full paper at http://www.mark-it.co.uk/text/asc.doc

It is becoming a standard for other systems too – ie software packages that are not data tabulators. MARSC http://www.marsc.co.uk/] system for extracting sample from CRM databases uses triple-s as the format for exporting the interviewer contact file. Nebu http://www.nebu.com export data and descriptions from their CATI and web-based interviewing package (DUB Interviewer) into triple-s.

As some software houses build analysers that work directly off triple-s files (to include hole-counts, tabulators, data/edit checkers etc), we can see the time for much great harmonisation and cooperation between all the major players. A link between dimensions and triple–s will open up the door to the world.

For further information contact the author at mark@mark-it.co.uk or visit his web site at www.mark-it.co.uk

Other References

Bethlehem; J -1999 - The Routing Structure of Questionnaires:
http://neon.vb.cbs.nl/rsm/tadeq
The IMS Project - http://www.imsproject.org/
The IDL Language -http://sda.berkeley.edu/idoc
Raosoft Project: http://www.raosoft.com/xml/oefs.html
W3C 2001 -XFORMS 1.0 - W3C Working Draft - http://www.w3.org/TR/xforms