The Challenge of the Internet

Proceedings of the ASC International Conference on

Survey Research Methods

Latimer Conference Centre, Chesham, UK,
11 – 12 May 2001

Editors:

Andrew Westlake
Survey & Statistical Computing, for ASC

Wendy Sykes
Social Research Association

Tony Manners
Office for National Statistics

Malcolm Rigg
BMRB, for MRS

Association for Survey Computing

Association for Survey Computing

PO Box 60, Chesham, Bucks, UK, HP5 3QH
Phone & Fax: +44 1494 793 033
E-Mail: Admin@asc.org.uk

ISBN: 0 9521682 7 8

British Library Cataloguing in Publication Data.
A catalogue record for this book is available from the British Library.

Contents

Designing on-line studies to maximum advantage

Sampling and Instrumentation Issues

Dissemination of Statistical Information

Integration through Software and Metadata

Preface

In 1998 the ASC ran a 2-day conference (entitled New Methods for Survey Research) at Chilworth Manor, Southampton, as a satellite meeting for Compstat (which was held in Bristol that year). Although the conference was relatively small (about 100 participants) the format, of a residential event and no parallel sessions, was very successful, allowing the participants lots of time to get to know each other and the speakers, and to explore issues raised by the presentations.

Because of this success, the ASC decided to include this type of event in its programme of conferences, continuing to focus on Survey Research Methods. Preparation for the 2001 event started soon after the 3-day, multi-stream, international conference held in Edinburgh in 1999. Building on the success of several previous one-day conferences, we decided that this event should concentrate on the Internet, looking not only at the new opportunities that it presents, but also at issues that cause concern.

As with the Chilworth event, ASC joined forces with related organisations for the planning of the scientific programme, the Social Research Association (SRA), the Office for National Statistics (ONS) and the Market Research Society (MRS). ASC took responsibility for the practical arrangements.

The Scientific Programme Committee decided to again organise the conference as four half–day sessions over two days, covering different (though related) topics. Each starts with an invited (keynote) presentation setting out the main ideas and issues involved, followed by three contributed presentations on the practical issues and the benefits associated with the topic. Each organisation has taken responsibility for one session, finding the keynote speaker, taking the lead in choosing, reviewing and editing the contributions, and chairing the session.

The following topics were chosen, falling naturally into two pairs for the two days.

	Topic	**Organiser**
Day 1:		
Session 1	**Designing on-line studies to maximum advantage**	Malcolm Rigg, MRS
Session 2	**Sampling and Instrumentation Issues**	Wendy Sykes, SRA
Day 2:		
Session 3	**Dissemination of Statistical Information**	Tony Manners, ONS
Session 4	**Integration through Software and Metadata**	Andrew Westlake, ASC

The first half of the book is concerned with data collection and quality on the Internet. In the first section, Ray Poynter provides the keynote paper, reviewing good practice for Internet surveys. This is followed by contributed papers reporting experiences and giving advice on how to get the best from Internet surveys.

Mick Couper then discusses some of the pitfalls that await the unwary web interviewer, to kick off a section about issues associated with mode effects and sample design. The contributed papers in this section discuss problems that have been experienced, particularly those related to response rates and non-response.

The second half of the book is concerned with the way in which the Internet and related new technologies have changed the options for disseminating information from and about surveys, and for the construction of survey processing and other statistical systems. Len Cook discusses the plans that are being developed by the Office for National Statistics for dissemination of their information. The first contributed paper reports experiences when applying this approach for a specific government survey, Rory MacNeill then reports on a web system that supports survey administration as well as the reporting of results, and the paper by Brannen and Lamb describes the Mission project, which is developing a statistical reporting system that operates across multiple databases with the possibility of definitions that require harmonisation as the information is analysed.

The final session is about models and metadata for building and using survey software. Although the greatest direct impact of these ideas is on the software developers, the motivation is to provide better facilities for those using survey software for design, data collection, analysis and dissemination. The keynote paper is by Chris Nelson, who reviews the standardisation processes used by the Object Management Group (OMG) and reports how these have impacted on the IQML project for ‘intelligent questionnaires’. The next two papers address related issues in the use of metadata to simplify intercommunication and interoperability between different statistical processes and different statistical organisations. Finally, Peter Andrews from SPSS reports on their proposals for a standard data model for survey data.

We have enjoyed the task of building the conference programme and editing this volume, and we trust that you will find the result enjoyable and informative.

Andrew Westlake, Wendy Sykes, Tony Manners, Malcolm Rigg

References

Westlake *et al* (Eds) (1998): *New Methods for Survey Research*. Association for Survey Computing, ISBN 0 9521682 3 5

Acknowledgements

The organisation of an international conference is always a team effort, and its success brings credit to all those involved. The editors were assisted on the SPC by two members of the main ASC committee, John Francis and Randy Banks. The general organisation effort was shared among the whole ASC committee, but special thanks are due to Diana and Steve Elder, for their administrative and financial skills, and to the ASC Chair, John Francis, for tirelessly chasing details and keeping everyone in order.

About the ASC

The Association for Survey Computing (ASC), originally known as the Study Group on Computers in Survey Analysis (SGCSA), was formed in 1971 in order to improve knowledge of good practice in survey computing and to disseminate information on techniques and survey software.

The ASC is a non-profit organisation, affiliated to the British Computer Society and the International Association for Statistical Computing. It has a wide-ranging membership at both individual and corporate levels, and has close working links with the Royal Statistical Society and the Market Research Society. Although based in the United Kingdom, it has a growing international membership.

Today its aims are:

- To act as a forum for the various disciplines within survey research and statistical computing
- To inform members of the latest software packages and techniques
- To organise regular conferences and workshops on key topics within the industry
- To disseminate information via its web site and publications
- To catalogue current software systems

The Association has begun sponsoring students so that they can attend courses in our discipline. It has paid the registration fees and accommodation expenses for two students at the IASC-IASS Summer School in Capri on the 20th-30th June 2001. The students have been asked to write up their experiences of the course for publication on the website.

The Association tries to keep people up to date with the role of computers in all aspects of the survey research process. In the past this was largely limited to collecting and analysing survey data. Over the 30 years of the Association's existence the role has widened into the design and formulation aspects of surveys and the presentation process. Now even more possibilities are emerging in the fields of integration, dissemination and communication of survey data and survey results.

The ASC organises four series of conferences:

- Three day multi-stream international conferences that cover all aspects of survey computing. The first one was in Bristol in 1992, then London in 1996 and Edinburgh in 1999.

- Two day single stream residential international conferences that have time to investigate the latest practices in survey computing. The first one was in Southampton in 1998 followed by Latimer in 2001.
- For many years the ASC has organised bi-annual one-day conferences that concentrate on a single topic, methodology or medium. These have been held in London and Edinburgh.
- We have had an occasional series of more specialised workshops that contrast and compare solutions as offered by different practitioners. These have always been London based.

In 1999 the ASC decided to use the Web as its primary medium for publication. For example, its major publication, the Register of Software for Statistical and Survey Analysis (which lists details of all relevant packages known to the Association) is now available from these pages. We continue to produce paper publications, including the Software Register and our occasional journal 'Survey Computing', but using the web site as the major source of material. The ASC has also published the Proceedings of the International Conferences, as well as more specialised monographs based on one-day events.

The activities of the ASC are organised by a committee of volunteers, supported by a part-time Administrator. The current ASC Committee is:

President:	Beverley Charles Rowe
Chair:	John Francis
Vice-Chair:	Randy Banks
Treasurer:	Steve Elder
Secretary:	Charles Clunies-Ross
Administrator:	Diana Elder

Committee: Thomas Brennan, Crawford Christie, Suzanne Evans, Hugh Gentleman, Laurance Gerrard, Raz Kahn, Tim Macer, Jennifer Waterton, Andrew Westlake, Peter Wills

Membership is inexpensive and members are entitled to a free copy of the Software Register and Survey Computing, and a reduction on other publications such as collected conference papers. Non-members are required to join in order to attend conferences, but membership is open to all interested in the subject area. At present, the annual membership fee is £25 for individuals and £75 for corporate membership. 3 year individual membership is £65. Corporate members may nominate up to 4 individuals as representatives (see the constitution), or more on payment of a small additional fee (£7.50 each). Please contact the ASC Administrator (Admin@asc.org.uk) for more details.

Details of ASC publications and activities, both past and future, are available on the world-wide-web, at http://www.asc.org.uk.

John Francis
Chair, ASC

Social
Research
Association

About the SRA

The Social Research Association was founded in 1978 as a forum for all those working in the field of social research. The Association's principal aims are:

- to provide a forum for discussion and communication about social research activity in all areas of employment;
- to encourage the development of social research methodology, standards of work and codes of practice;
- to review and monitor the organisation and funding of social research
- to promote the development of training and career structures for social researchers
- to encourage the use of social research for formulating and monitoring social policy

These aims are promoted by an elected Executive Committee, which carries out most of its work through sub-committees concerned with **events, training, public relations, membership** and **publications**. A quarterly Newsletter keeps members in touch with the work of these committees, as well as providing them with the opportunity to write about their work, social research issues and publications.

Our website (www.the-sra.org.uk) contains a wide range of subjects including information about careers, job adverts, SRA training days, seminars and conferences, social research articles, our advertising rates as well as links to other useful websites connected with social research.

About the ONS

The Office for National Statistics (ONS) is the government agency responsible for compiling, analysing and disseminating many of the United Kingdom's economic, social and demographic statistics, including the retail prices index, trade figures and labour market data, as well as the periodic census of the population and health statistics. It carries out most of the government's major household surveys, including the Labour Force Survey, the General Household Survey and the Expenditure and Food Survey. The Director of ONS is also the National Statistician and the Registrar General for England and Wales, and the agency administers the statutory registration of births, marriages and deaths there.

National Statistics is the official source for authoritative, accurate and relevant information on the UK's economy and society. It brings together a vast range of statistics overseen by the National Statistician.

From the nation's wealth to the state of its health, National Statistics data say something about every aspect of our lives. They tell us what we earn and how we spend our money. They highlight trends in marriage, divorce and births. They help medical research and improve our understanding of society.

Everyone plays a part in National Statistics. Information from census returns, and social and health surveys, is used to help to allocate resources for schools, hospitals and other amenities. The details we supply allow the government to manage the economy and enable industry to plan for the future.

National Statistics data are produced to high professional standards set out in the National Statistics Code of Practice. They undergo regular quality assurance reviews to ensure that they meet customer needs. They are produced free from political interference.

About the MRS

The Market Research Society is the UK's professional body for individuals involved in Market and Social Research. It is the largest body of its kind with over 8,000 members working in organisations currently undertaking market and social research in the UK and overseas.

What we do?

- We cater for our membership's needs. We have a diverse membership of individual researchers within agencies, independent consultancies, client-side organisations, and the academic community at all levels of seniority and in all job functions
- We are the leading provider of professional training courses "h We are the official awarding body for qualifications in market research
- We operate a Code of Conduct approved throughout Europe providing an ethical framework for conducting research
- We are the 'voice' of the profession in our media relations and public affairs activities on behalf of research practitioners
- We aim to achieve the most favourable climate of opinion and legislative environment for research

The following are a range of MRS activities:

- The Annual Conference and Trade Fair of the Market Research Society, the Society's flagship event; a forum on developments of the industry and its visionary future
- Research magazine - the best independent source of news on the industry worldwide (available on subscription)
- MRScene - a journal keeping MRS members in the picture
- Research Buyers Guide - a directory of organisations and consultants based in the UK and Ireland offering market and social research and related services (available on subscription)
- 70 different training courses geared towards enhancing skills and professional competence

Designing on-line studies to maximum advantage

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

A Guide to Best Practice in Online Quantitative Research

Ray Poynter

Abstract

In the year 2000, about 5% of the Market Research conducted in Western Europe and the US was conducted using the medium of the Internet. This is expected to rise to about 8% this year, and reach 50% by 2004. This change is being accompanied by many other changes; changes in business models, changes in legislation, changes in respondent resistance, and changes driven by lifestyle and technology.

The paper sets some out thoughts on Best Practice for Internet interviewing, in the context of the year 2001. These thoughts and suggestions have progressed from the ones written by the paper's author in the 1998 ESOMAR Handbook of Market and Opinion Research. In this field we tend to say that one Internet year is seven in the real world - so we all had better treat Internet advice as work in progress!

Keywords

Best Practice, Online, Paradata, Sampling, Ethics

1. Introduction

In the year 2000, about 5% of the Market Research conducted in Western Europe and the US was conducted using the medium of the Internet. This is expected to rise to about 8% this year, and reach 50% by 2004. This modality change is being accompanied by many other changes; changes in business models, changes in legislation, changes in respondent resistance, and changes driven by lifestyle and technology.

The research industry has been busy reviewing its procedures and processes, and in doing so it has had to confront some long neglected truths. For example, conventional market research has not been able to collect data based on random probability samples for many years, but with the rise of the Internet it has had to confront the problem and be clear about how it will handle it. Our conventional questionnaires have been getting longer over the years, with an increasing number of questions that baffle or bemuse the respondents. With the Internet, questionnaires have to be short, clear, and to the point, or they are not answered. In these two ways, and in others, the Internet is obliging researchers to address long-standing problems and define processes.

This paper sets out some thoughts on Best Practice for Internet interviewing, in the context of the year 2001. These thoughts and suggestions have progressed from the ones written by the paper's author in the 1998 ESOMAR Handbook of Market and Opinion Research. In this field we tend to say that one Internet year is seven in the real world - so we all had better treat all Internet advice as work in progress!

2. Judging the Web

Online zealots adopt every web option, no matter how unsound. Research reactionaries reject all online options, often quoting reservations that are minor compared with the lapses they accept daily offline.

This paper makes the proposition that we should judge online research using exactly the same standards as we judge offline approaches. With all offline techniques we have to accept imperfections in order to be pragmatic. When we review online techniques we should be aware of their imperfections, and contrast these imperfections with those of other approaches. The key test for an offline technique is, does it improve decision-making? When we consider an online approach, we should be asking the same question.

This paper urges that we should listen to the words of the Roman, Marcus Aurelius Antonius, "Never let the future disturb you. You will meet it with the same weapons of reason which today arm you against the present".

3. Self-Completion

The biggest difference between Internet research and most offline research is not the presence of the Internet; it is the absence of the interviewer. In face-to-face, phone (CATI), and most CAPI interviewing the interviewer plays a vital role. The interviewer helps present the questionnaire, to persuade the respondent to do the survey, and to cajole/encourage the respondent into finishing the survey. If there are glitches in the interview or if the instructions are ambiguous, the interviewer will often help improve the communication and the process.

The respondent's interaction with the interviewer can itself be a rewarding process, enhancing the standing of the interview, and making co-operation with future research more likely.

With the Internet the questionnaire is all there is. If the instructions are not clear the respondent will make mistakes or even abandon the interview. If the interview is too long or too boring, respondents will tend to abandon the interview.

The key things to get right for self-completion are:

- Clear instructions
- Engaging design
- Error-free questionnaires
- Short interviews
- Treating the respondent with respect

Clear Instructions

Clear instructions start with the initial contact. The only thing we really know about the respondent's screen is about the top-left of their screen. Different monitors, browsers, font settings, and configurations will mean that an interview that fits nicely on your screen may not fit on the respondent's screen.

If you want people to scroll down – say so. If people can use the BACK button, say so. If people should not use the browser back arrow – say so. Do not assume, it will make an ASS of U and ME.

Engaging design

People expect a high standard of design on the Web. If your interview is tacky and unattractive it will hold on to fewer respondents. However, this does not mean that you can squander bandwidth on rich media, if respondents have to wait for your wonderful effects to download they will usually abandon the interview.

Error free questionnaires

As well as being engaging the interview must be error free! There is no interviewer to sort out any problems. A respondent who encounters an error is likely to simply abandon the interview – and be less likely to co-operate in the future.

This means testing the interview on different browsers, with different settings.

Short interviews

On the Net, interviews need to be shorter. This means reviewing our practices and stock approaches. As a guide the following table shows a consensus view of the target lengths for different types of interviews:

Pop-up/intercept	5 minutes
Ad hoc, modest incentive	5 to 15 minutes
Panel, normal	10 to 20 minutes
Employee research	10 to 20 minutes
Specialist panel/large incentive	10 to 30 minutes

It is possible to ask longer surveys, but you will need larger incentives and you will normally face lower response rates.

Treating the respondent with respect

Many of the comments thrown up by traditional interviewing systems were designed by programmers to remind CATI interviewers of procedures and to highlight errors. For example, one leading software demo prompts a missing click in a grid with "No that's not right, try again". Prompts, in particular those asking for the respondent to correct a response should be clear, helpful, and encouraging.

4. Sample

The main concern about online interviewing relates to sampling. Key questions include:

- How representative is the sample of target population?
- Who is really at the other end?
- How seriously are interviews completed?
- Are respondents self-selecting?
- Are we talking to 'professional respondents'?

A presidential sample warning

In 1936, in the USA, the Literary Digest conducted a poll in order to predict the winner of that year's presidential election. The Literary Digest had conducted similar and successful polls on several of the previous elections. They mailed out 10 million questionnaires and received 2.3 million replies. They printed their story and predicted a 3:2 victory for Alf Landon. In fact, FD Roosevelt won all but 2 states!

What went wrong? The majority of the 10 million recipients of the Literary Digest mailing were car owners, magazine subscribers, and telephone owners. During the recession in the USA these were not an unbiased group of people – they disproportionately favoured the white-collar candidate over the blue-collar candidate!

This case serves to remind us that sampling can matter, even when the sample is very, very, large.

A new presidential example

In the recent US Presidential election in the US, there were a large number of polls, some predicting a win for Gore, some for Bush, and some calling it too close to call. Amongst the most accurate of all the polls was the online survey from Harris Interactive – who called it too close to call, with good estimates of both candidates' percentages.

By using online polling successfully (including Senate and Gubernatorial races) Harris Interactive has done a service to all of us by showing that online sampling does not have to repeat the errors of the long demised Literary Digest.

Pragmatic approaches

Approaches are being developed online to handle the sampling problems of general research, and these approaches, in most cases, mirror what is happening offline. The most common approaches are:

Ignore the problem

Systems such as www.insightexpress.com allow the user to access a sample, design a simple questionnaire, pay by credit card, and start receiving answers. At its most basic there is no way to assess the population, the response rate, or any other statistical feature. This approach is likely to lead to problems!

Recognise the problem, and live with it

A good example of this approach is brand tracking. The key measurements do not immediately relate to external facts. For example brand awareness and positive image do not have a precise mathematical link to net sales. If a tracking sample is constructed to be stable over time, then the rises and falls it measures are likely to represent the rises and falls in the whole population. These rises and falls can be referenced against external facts such as sales. Tests by companies such as Millward Brown have shown this to be the case.

Ameliorate the problem with Quotas

This approach is similar to the pragmatic approach taken with most offline research. If researchers ensure that they have quotas that collect a sufficient number of each of the key socio-economic groups (or any other relevant quotas), then problems will generally be minimised. The only occasion when this will not help is when the Internet is highly associated with a source of bias (see next section on topic and medium bias interaction).

This approach is very well suited to target sources such as access panels and to large sources such as AOL's Opinion Place in the US.

Advanced sampling approaches

There are statistical techniques other than random probability testing. For example in epidemiology it is often necessary to live with data where patients make choices about their treatment that ruin random probability designs. Harris Interactive in the US has adapted one such approach, where they have used a method based on the propensity of an individual to be online. This has allowed them to weight the data to make it more likely to represent the non-virtual world. This approach is more robust than the quota approach, but will still fail if the medium is highly associated with the topic bias.

It should be noted that Harris's approach is only suitable in cases where there are large online panels and detailed offline information to match it. This is unlikely to be the case outside the US.

Norms based approaches

Many companies are developing norms that allow them to interpret the data they collect from the web in ways that have direct relevance for the non-virtual world. The typical approach is to obtain a stable set of data collection approaches and then to calibrate these to the non-virtual world and to previous data.

Examples of this approach include ACNielsen/Bases and their concept testing in the US and Millward Brown's advertising pre-testing with Link.

Modality bias, topic bias and sample bias interaction

This section addresses the phenomenon that the wrong sample often gives the right answer. Since the advent of Internet-based research there have been many side-by-side surveys. Most of these have found that the conclusions and recommendations from the online research have closely mirrored the offline – even when the samples have been quite different.

The wrong sample will tend to give the right answers if the bias in the sample is independent of the views held about the topic. For example if we were to ask whether people were left-handed or righthanded we could ask most biased samples and still get a fair result. This is because the distribution of left-handedness is largely independent of sample effects (be they sex, social-class, or online access).

In terms of the Internet this implies that the wrong sample will provide the wrong answers when either of the following is true -

- If the views about the topic differ in ways that mirror differences in online access. In the UK this means that if views about the topic are differentiated by age or social class we would expect the online research to provide misleading answers. In these cases, controlling the sample variances with quotas will tend to be helpful.
- If the views about the topic differ between online people and offline people. For example, views about CDs, their prices, the purchase locations, and their availability are likely to be markedly different between online and offline people. Quotas will not help alleviate this problem.

5. Opt-in, Opt-out, and the future

The current debate about contacting respondents is centred on the concept of permission – a term that has been championed by Seth Godin in his book "Permission Marketing". As a result of legislation, ISP rules, respondent resistance, and current best practice the drive is towards contacting people who have given permission and who anticipate being contacted.

The US tends to favour an approach called opt-out. With this approach the researcher ensures that there is a simple and reliable mechanism for somebody to say 'do not contact me in the future'. The term is also used to cover registration forms that have a box that has to be ticked to say, "do not contact me".

Europe, by contrast, tends to favour opt-in, and certainly the European Union appears to be minded to legislate in favour of opt-in. The key ingredient of an opt-in scheme is that people have to actively say 'contact me'. For example on a registration box there may be a box that says "I authorise you to contact me for …..". Sometimes companies pre-tick the opt-in box; this blurs the distinction between opt-in and opt-out. In the very worst cases this box is positioned off the main viewing area!

It should be noted that even where an opt-in scheme has resulted in the permission, there should still be a simple and reliable method of withdrawing permission.

Companies that operate a double opt-in system provide the highest standard of permission. With the double opt-in, the first stage is that somebody agrees to be contacted and supplies his or her email address. However, before that person is deemed to have given permission they are emailed and asked to confirm their permission. This system is used by most of the leading panel companies.

6. Contacting the respondents

This section outlines the most common ways of bringing respondents and the questionnaire together.

Pop-Up

The online intercept is a very popular form of recruiting respondents. A piece of script is used to pick every nth visitor to a site, or page, or feature. If a respondent is selected then the system 'pops up' a new browser window in front of the window the respondent was viewing. This new window contains an invitation to do the survey, and one or more methods of electing not to do the survey.

Usually, the system will serve the respondent's machine with a cookie to ensure that they are not invited again. This means people who clear out their cookies are prone to be invited more often, people with more than one machine may be invited twice, and amongst people who share a computer only one is likely to be selected.

Pop-ups are used extensively to assess websites and services (eg Millward Brown's Audience Audit). These interviews are used to find out who is visiting a website, what are they looking for, what their experiences are, etc. Pop-ups are also used extensively to create controlled experiments in testing alternative online experiences. For example Millward Brown's Brand Impact tests online advertising by serving half of a sample a test advert, and the other half a control advert. After the respondents have been exposed to either the Test or Control advert they are then presented with a pop-up questionnaire. This sort of approach ensures that the only difference between the Test and Control samples is the online stimuli.

The keys to successful pop-ups are:

- short interviews;
- appealing layout;
- quality of invitation;
- relevant topic of research.

There is disagreement about whether it is appropriate to offer an incentive for pop-up interviews. Pete Comley of Virtual Surveys has argued that offering incentives attracts fraudulent respondents, and encourages people to delete their cookies so they will be invited more than once.

Response rates for pop-ups vary. Fletcher Research has reported rates as varying from 2% to 15%. By contrast Virtual Surveys have published their data showing an average of 25%.

Hijack

The Hijack is similar to the Pop-up, but is more intrusive. If a respondent is selected for an interview they are re-routed to an interview page, instead of the page they had requested. The respondent should be offered a decline button, which will take them to the page they were expecting to go to.

In the UK, the BBC have used these Hijack interviews with great success (response rates of about 80%, no incentives, and few complaints). However, this probably reflects the positive relationship the BBC has with its audience.

In general the Hijack approach is not recommended, as it is too intrusive, and would probably increase future respondent resistance.

Invitation tags

Websites abide with requests for people to take surveys. These vary from customer satisfaction to topic-centred research. The format of the invitations varies from simple text to graphical features.

The people who choose to do these surveys are entirely self-selecting and are often differently motivated from the average site-user. However, this sort of recruitment can be useful in monitoring changes in satisfaction and in collecting input about topics of current interest. But this approach should not be used instead of more objective techniques.

Banners

Banner ads can and have been used to invite people to take a survey, usually in conjunction with an incentive. Traditionally these ads were for specific surveys, particularly amongst hard to recruit groups. The main benefit of this form of recruiting is that the advert can be placed on websites that specifically target certain types of users, eg finance, auto, parents of young children, etc. The problems with this approach are low response rates (less than 0.5% typically), high cost (because of the low response rates), self-selection by the respondent in terms of wanting to do an interview and the topic of the interview.

More recently banners have become popular as technique to attract people to go to opinion centres such as AOL's Opinion place. A possible benefit of these ads is that they do not tell the respondent in advance what the topic of the research is going to be.

Opinion Centres

The idea of an opinion centre has been popularised by AOL/DMS and their OpinionPlace (www.opinionplace.com). By utilising the massive presence of AOL, and its partners, the system attracts people to visit the Opinion Place. People who visit the site decide to do a survey and they answer a few simple questions. They are then allocated to a survey. The respondent chooses when they want to do a survey (up to one every two-weeks), but they do not choose the topic – that is chosen by the system.

Because of the large numbers of people available in the US through OpinionPlace this approach is very suitable in cases where quotas are going to be used to control the sample.

Panels

Companies such as Harris Interactive, LightSpeed, and Greenfield have large online access panels in the US, and are rolling these out across Europe. These panels comprise of people who have agreed to be contacted and about whom the panel companies has stored some profile data. This means groups can be targeted and quotas filled to ensure that samples remain stable over time and in order to match samples to specified targets.

There tends to be a debate as to what really constitutes a panel. Some of the larger panels are closer to databases of permission based email addresses. Some of the smaller, more focused panels have few members but extensive profilers. For example the Millward Brown IntelliQuest Tech Panel has over 20,000 members and a profiler of about 500 questions.

Normally, the communication with the respondent is dealt with by the panel company, and in many cases the interview must be written and administered by the panel company. Typical response rates for a panel are 30% to 40%.

Email lists

Companies such as SSI (Survey Sampling Inc) maintain large databases of email addresses from people who have given permission for contact. Often these permissions will have been given incidentally whilst signing up to receive some other service. Most panel members know that they are a member of a panel. Most people on an email list do not know they are on that specific list, as they will tend to be on many lists.

Typical response rates from high quality email lists are 2% to 8%. Lists can be useful in providing stable samples for tracking studies. In order to use them with quotas, it will usually be necessary to invite a very large number of people. If a list has a response rate of 5% and only 20% of people are likely to screen in, then you need to invite 100 people for each completed interview.

Private Lists

In addition to commercial email lists, you or your client may have email contact lists (for example customer lists). The key issues to consider here are:

- Permission, do you have the consent of these people to use their information to contact them?
- Can you identify children?
- Do you have an unsubscribe method? – you need one!
- Are the email addresses transcribed correctly?
- Are the email addresses up to date? - people change their address quite often.

Phone invitation

This is a relatively specialised option, mostly suitable for B2B research, typically recruiting to panels. The technique has also been successfully used to recruit geographically clustered consumer samples, eg for retail products. The phone interviewer contacts the respondent and elicits that they are qualified and are prepared to join the panel. The interviewer then talks them through the online registration process, reading them the simple URL over the phone. This approach is more expensive than most other approaches, but results in higher completion rates.

7. High Frequency Respondents

There has been increasing concern over 'professional respondents' (that is respondents who complete many surveys). In the UK there is evidence that the average housewife, who is interviewed for a telephone omnibus survey, does 11 surveys a year. Anecdotally we hear that 50% of all responses in the US come from just 4% of the population.

Although this problem exists in the offline world it can be much worse with online interviewing. This is because there is evidence that some people are seeking out surveys to complete. The are a large number of websites (such as www.surveys4money.com, www.doughstreet.com, and www.rewardsites.com) that tell surfers where to find surveys from which they can earn money or win prizes.

If you run a survey with a nice, attractive incentive and all people have to do is to go to a specific location to complete the survey; you will get some people who are there just for the incentive.

The issue of high frequency respondents is intensified by the growing reliance on panels and email-lists. The people on these lists are by definition high frequency respondents. Different panels have differing policies on how often they interview people, for AOL/Opinion Place and the Millward Brown IntelliQuest Tech Panel it is once every two weeks, for SSI it can be up to twice a week. However, none of the panel and email list companies can control how many other panels and lists their members are on.

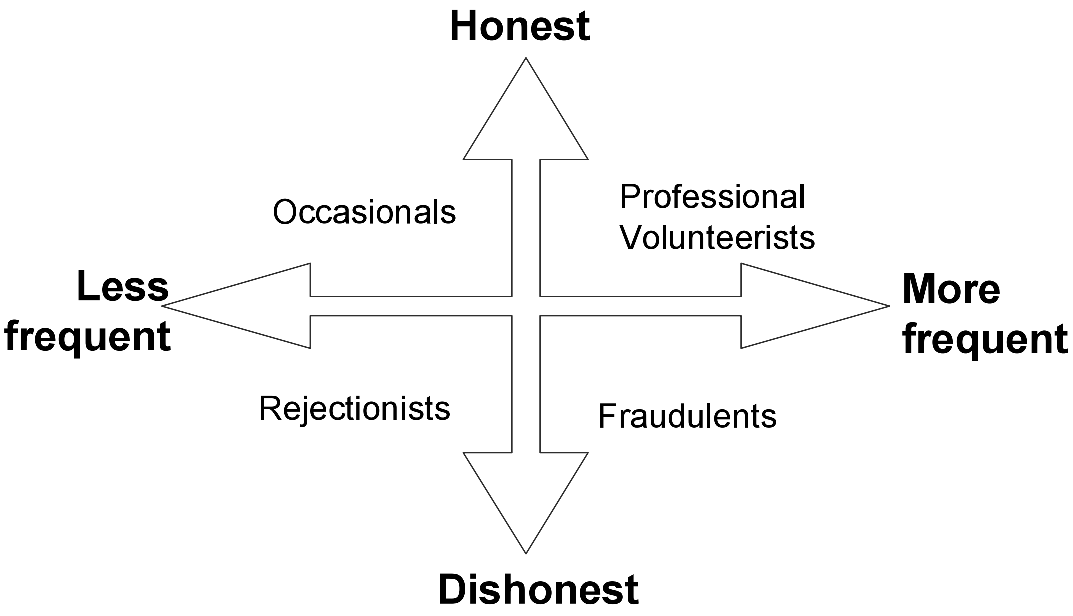
The people on these lists certainly seem to want to do a large number of surveys. In a survey conducted by SSI, almost 90% of those who replied wanted to do at least one survey a week, with over 60% wanting to do more than one per week. In a paper by Wilke, Lundy, Mustard, at Net Effects, Dublin 2000, they investigate the ACNielsen/BASES panel and found that 46% of people who were doing more than 20 tests a year felt that they wanted to do surveys more frequently.

Frequency and Honesty

The issue of frequency, and of people seeking out surveys so they can obtain the rewards raises two concerns.

- Do people who respond to more surveys become more sensitised to certain issues and therefore less representative of the rest of the population?
- Do people tell the truth, particularly if they are being motivated by incentives?

The diagram below spells out the way frequency and honesty interplay:



In this categorisation of respondents the greatest worry is the group termed Fraudulents, this people complete a large number of surveys and tend to give incorrect answers, either to maximise their earnings or out of malice. When researchers have looked for these people, for example by searching panels for inconsistent responses, they appear to be few in number.

The Professionals are high-frequency respondents who are motivated by incentives, but who tend to be truthful. Volunteerists are high-frequency respondents who are motivated either by a desire to help, or because they find surveys fun! In a study carried out by SSI in March 2000, they found about 50% were Professionals and 50% were Volunteerists.

The Rejectionists are people who do not want to do surveys. Some people are Rejectionists for all media. Others would reject an online interview, but might do a store exit survey, for example. There is evidence, for example from the MRS Respondent Interviewer Interface work, that the number of Rejectionists is steadily growing. There is also some anecdotal evidence that when these people are trapped into doing interviews (by a persuasive interviewer in the offline world, or in order to receive a service online) they are likely to be less honest.

The Occasionals are the people researchers would like to reach more often. If asked to do a survey at the right time, on the right topic, in the right way, they will do the survey. These people are quite rare in panels and email-lists, but they can be reached by techniques such as pop-ups.

More Research Needed

More research is needed on the impact of high-frequency respondents on the research they are involved in. Do high-frequency respondents provide different conclusions and recommendations? Are there important differences between Professionals and Volunteerists? What is the best way to spot Fraudulents?

Wherever possible add a question to your survey to ask "In the last six months how many surveys have you taken part in?" Add another asking all those for whom the answer to the previous question is more than one "What is the most important reason you complete surveys? Opportunity to influence decision makes, Rewards and Prizes, Fun, Interesting topics, To help the interviewer."

Explore new ways of identifying Fraudulents, perhaps incorporating Paradata in the process.

8. Response rates

The higher the response rate the lower the sample cost and the higher the comfort factor in terms of non-response bias. The following headings cover steps that can be taken to increase response rates.

Self-completion issues

All of the issues addressed earlier in the paper in the self-completion will impact response rates.

Interview Length

This was covered earlier, but it is worth repeating. Keep the interviews as short as possible. Consider spreading questions across more than one questionnaire.

Ask the right people, the right way

Scatter gun approaches to sample recruitment tend to result in low response rates, and more annoyance amongst the recipients of those requests. The more a researcher can target the requests to people who are relevant to the topic, and to people who have given permission to be contacted, the higher responses will be.

If people are contacted by email then make the process clear. If they should visit a web site first tell them, if they are going to need some information about their company, say so. Tell people how long the interview will be - do not let them discover it as the interview drags on. Informing respondents of the expected length of the interview is an MRS requirement.

Incentives

Incentives are increasingly common and necessary on the Web. Given that many respondents pay for their connect time to the Internet (usually to a telecom company) it is also an ethical issue.

The consensus is that for most forms of interview, incentives increase response rates. However, small incentives and prizes that are not of universal appeal may produce no beneficial results.

The first decision the researcher needs to make is whether to go for a one-to-one incentive (where everybody is rewarded), or for some form of lottery or prize draw. Prize draws are popular with researchers since they are easy to estimate and they cap the expenditure at a particular cost, irrespective of the number of replies.

Some prefer one-to-one incentives because they believe that they result in a higher response rate. At present this point of view has to stay a belief, since the data that has been published is somewhat contradictory.

Methods of incentivising

There are a variety of methods of incentivising respondents.

- Cash. Cash is very popular as a prize draw incentive. However care needs to be taken to make sure the incentive is large enough to motivate additional responses. Small incentives, eg £100, tend not to be motivating. Best practice at the moment is to try and have an attractive first prize and several minor prizes, eg £1000, plus five prizes of £100. Cash does not work well as a one-to-one incentive, because of the cost of administrating it, unless working with an online panel with a remuneration system already set up.
- Amazon vouchers. These are a popular form of one-to-one incentive, mainly because they are easy to administer. All that is needed is to email a string of digits. They are also easy to redeem and of widespread appeal. However, not everybody is motivated by Amazon vouchers.
- Points. There is a range of online points systems, such as Beenz. Surfers can collect points by visiting websites, making purchases, and by taking part in surveys. The great advantage of these systems is that they are cheap to administer and therefore support one-to-one incentives. The main disadvantage is that these systems are very motivating for those who take part, but of little interest to non-subscribers. One variant of this system, which does appear to work well, is the AOL American Airlines' Advantage reward scheme. These airmiles can be used in the conventional way and to make a variety of purchases.

- Results. Many researchers report success, in B2B research, when they offer an abridged copy of the results as an incentive, or as part of the incentive.
- Prizes or gifts. Care must be taken with prizes or gifts. If the prize is large enough and of sufficient appeal it can be motivating. For example Millward Brown IntelliQuest's Tech Panel offers a prize of a Toyota Lexus for membership of the panel. However, if the prize is not motivating it can actually reduce response rates. For example, a paper at an IQPC conference in London showed a case where an incentive of a chance to win two free tickets to the London Eye (a large Ferris wheel), reduced response rates!

Disclose sponsor

Many respondents say that one of the reasons they take part in surveys is to help companies produce better products. These people will be more motivated if they know whom the research is for. This approach is not always relevant, but it should be considered each time.

Optimise questionnaire

Many of the earlier points in this section relate to getting as many suitable people as possible to start the survey. The final step in maximising the response rate is ensuring that as many people as possible complete the interview.

The first element in this process relates to points already covered in terms of making sure the design is professional, that the questionnaire is error-free, that it runs on a range of browsers and settings, and that the respondent is treated with respect.

One of the newest topic areas in online research is Paradata. Paradata is data about the process. In online research we have lost the interviewer, but we have gained the ability to monitor the process. We can measure how long interviews took, we can measure how many respondents abandon at each stage, we can review the browsers the users have, and the settings they are using. This mass of information is being used to increase the level of knowledge about our survey instruments.

Paradata should be used to optimise questionnaires. Using Paradata approaches you can monitor the questions that are taking too long, the questions that are associated with corrections, and the questions that are associated with abandoned interviews.

Across a range of studies Paradata can allow you to build up a set of rules to help you optimise studies. Other researchers such as Andrew Jeavons have published their findings. These show us that grids are particularly associated with both errors and abandoned interviews. Questions where people are supposed to type in numbers that add up to some specific number are also highly associated with errors and abandoned interviews.

Research by Millward Brown has shown that verbatims placed near the beginning of an interview, particularly if they are 'touchy feely' will result in an increased number of abandoned interviews. Work by Pete Comley has shown that forcing completion will also result in more abandoned interviews.

9. Ethics

Ethics is a key issue for Internet research. Key topics include email, data protection, surveillance, tracking, and children. Over the past few years, practices, guidelines, and laws have all been in a state of flux. This will continue in the future, meaning that researchers will need to adapt to the process of adapting.

Data Protection and Data Privacy Laws

The key directive in this area is the 1995 European Directive, implemented in the UK as the 1998 Data Protection Act. However, there are other laws and directives covering distance selling, e-commerce, and telecomms. In addition, there is a stream of interpretations from Europe, the Data Protection Commissioner and courts.

Since 30th January 2001, the Data Protection Commissioner has been known as the Office of Information Commissioner and is also responsible for Freedom of Information.

The general trend in Europe is towards demanding opt-in, and towards banning unsolicited contact (by post, phone, or email). The big picture is that companies should be open and clear in their dealings and they should only collect and keep such data as they would reasonably need.

Outside of the research industry there is a consensus that the exceptions and dispensations that exist for market research will disappear over time, and the market research industry will increasingly be treated more like marketers. This view is not shared by most of the bodies representing market research organisations.

Best practice for researchers is:

- tell people what you are doing;
- ask people for permission;
- do only what you said you were going to do.

The rules mean that the process for exporting personal data outside of the European Union is quite onerous. If you plan to do this take special advice, and ensure that you let respondents know that their data will be stored or processed outside the EU.

More information is available from www.dataprotection.gov.uk.

ESOMAR/ARF Guideline Approach

The lead on ethics in Internet Research has been taken by ESOMAR, who have debated this, issued guidelines, and worked with other organisations since their Edinburgh Congress in 1997 and the first Net Effects conference in Paris in early 1998.

The latest version of the guidelines, officially the ESOMAR/ARF guidelines, can be downloaded from the ESOMAR website at www.esomar.nl. These guidelines are updated regularly, so it is a good idea to visit the site periodically to check for changes.

The guidelines do not seek to be prescriptive about the technology that should be used to avoid problems. ESOMAR recognised back in 1998 that the rate of change in the Internet was too fast to allow a detailed set of procedures to be drawn up. Therefore the approach that has been adopted is to highlight the normal research guidelines and to advocate that these be sensibly transferred to the web.

Privacy Policies

Your website should include a Privacy Policy, outlining what data you collect, what you do with it, what people can find out about the data held about them, and how they decline future contact.

Each online questionnaire should contain a short note about the key privacy aspects and a link to a full Privacy Policy. For example, the questionnaire should highlight the anonymous nature of the data, if cookies are being served, and whom they should contact with queries.

Children

The key requirement of the guidelines is that there should be prior consent. The key issues for researchers are -

- How do you recognise you are talking to a child? The best practice is to ask a question about age early in the survey.
- How are you going to obtain parental permission to interview the child? The most common and reliable way is to use panels, where the parental permission has already been obtained.

In an experiment conducted by Millward Brown, and approved by the UK's Market Research Society, children were contacted and asked to obtain their parent's permission. This permission was confirmed by the children typing the name of their parent into a field and providing a contact telephone number. When the parents were called back it was found that 30% of the permissions were completely false. In only 13% of cases had parents looked at the screen and made an informed view that this was OK for their children. This experiment confirms that runtime permission is not acceptable, with current techniques and technology.

Invisible processing

Invisible is a phrase that includes any data collection that the respondent is not immediately aware of, including the collecting of paradata.

The topic that has created the greatest interest to date is the cookie. The typical cookie is a time-dated text that is placed on the respondent's PC. One major use of cookies is to avoid re-sampling the same person. However, they are also used to customise settings and to track people across several visits. Best practice is to tell people if cookies are being served at the beginning of the interview and in the full privacy policy. The cookies should be dated to expire as soon as is consistent with the research.

Other invisible processing can include such things as discovering the ISP of the respondent, the speed of the connection, the Operating System, the browser, the screen definition, and the JavaScript settings. Best practice is to include a paragraph in the full Privacy Policy explaining about the sort of data that is collected and the use that it will be put to. It is not possible to detail everything that might be collected, since the researcher themselves will not always be aware of everything their system is collecting.

Spam

Spam is unsolicited email and is not permitted under the guidelines. It is also banned by most ISPs and is likely to be made illegal within the European Union in the near future.

In order to operate within the guidelines, agencies need to ensure that any list of email addresses they use has the appropriate permissions associated with it. Generally this means either using a list from a recognised vendor or asking the supplier for confirmation that permission exists.

Agencies also need to have a mechanism for recording anybody who asks to unsubscribe from that list, and for passing this information back to the original list holder. This is not common practice for agencies, but it is a requirement of the law.

Appendix Relevant Links

Information sites

www.cyberatlas.com

www.macer.co.uk
(includes references to many online data collection systems)

www.nua.ie

www.worldopinion.com

Online reporting

www.e-tabs.com

Online qualitative

www.itracks.com

www.parachat.com

www.vrroom.com

Sample sources

www.sampleanswers.com

www.surveysampling.com

Panel/Database companies

www.greenfield.com

www.harrisinteractive.com

www.insightexpress.com

www.lightspeedonlineresearch.com

www.surveysampling.com

www.techpanel.com

www.vrroom.com

Incentives

www.beenz.com

www.cybergold.com

www.globogift.com

www.igain.com

Online Measurement

www.mediametrix.com

www.netratings.com

Privacy

www.truste.com

www.esomar.nl

www.mail-abuse.org

www.imro.org

Anonymous Tracking

www.doubleclick.com

www.engage.com

References

Comley, Pete (2000): “Pop-up Surveys. What works, what doesn't work and what will work in the future”, Proceedings of the ESOMAR Worldwide Internet Conference Net Effects 3, Dublin, 2000

Godin, Seth (1999): Permission Marketing, Simon & Schuster, 1999

Jeavons, Andrew (1999): "Ethology and the Web. Observing Respondent Behaviour in Web Surveys". Proceedings of the ESOMAR Worldwide Internet Conference Net Effects 2 London 1999.

Jeavons, Andrew (2001): "Paradata: Concepts and Applications ". Proceedings of the ESOMAR Worldwide Internet Conference Net Effects 4 Barcelona 2001.

Wilke, Joseph, Lundy, Sheila, and Mustard, Donna, "Burning Out Internet Respondents. Avoiding the Mistakes of the Past", Proceedings of the ESOMAR Worldwide Internet Conference Net Effects 3, Dublin, 2000

Witt, Karlan and Poynter, Ray (1998): Research on the Internet, ESOMAR Handbook of Market and Opinion Research (pp 1085-1101), 1998

About the Author

Ray Poynter is the Director Europe of Millward Brown IntelliQuest and can be reached on ray.poynter@uk.millwardbrown.com.

Experiences of Qualitative Research on the Internet

Charlotte Cornish

1. Summary and Conclusions

After varied experiences of using the Internet as a qualitative research tool since 1997, we have muted enthusiasm about conducting qualitative research on the Internet. At the Future Foundation we will continue to use Internet-enabled techniques when off-line methods are unpractical or inefficient but not as a day-to-day replacement.

First, on the positive side – the major benefit for us has been being able to hold discussions between groups of people who would have been difficult to bring together by any other means. Geographically dispersed communities, for example, and other respondents who for one reason or another are not able to leave their home to attend a group - parents of new babies, carers of elderly relatives and so on.

Further, the Internet has also brought benefits to our clients. The use of moderated on-line groups and e-mail groups has meant that our clients have been able to watch the proceedings from their own desk. Clients have liaised with us while the discussion is in progress and have got a greater sense of involvement with the whole process. Indeed, we have found that clients' enthusiasm is often the major reason behind our choice of an on-line qualitative methodology in the first place.

Finally on the positive side, the Internet has saved us money. While recruitment and incentive costs are similar, there are no hosting fees, taxis or babysitters to pay for. Furthermore, transcription is not necessary and analysis is quicker because we have an exact record of the whole discussion.

There is a negative side, however. We have encountered three major problems. First, recruitment – not so much finding people who fit our criteria in the first place, with access to the Internet now accounting for over 50% of the population this is not a major problem. What we have had difficulties with is knowing how many to recruit to ensure a discussion group of a reasonable size – most common is a high drop out rate but on some occasions we have had too many people turn up! It seems to us that the main reasons for dropping out are the main reasons why people turn to the Internet in the first place eg. Time constraints and interruptions at home. But we have also had quite a few respondents drop out during proceedings due to problems with software or hardware.

Our second problem is depth of coverage. Our experience of this has been very mixed. In some cases our respondents' replies have been detailed and interesting but at other times very brief. The main reason for brief answers in our experience is unfamiliarity with the on-line research software and an inability to type quickly.

This brings us to our final problem, namely the difficulty of moderating on-line groups. The remote technology means that in most cases the group dynamic is diminished or lost entirely. This can result in a very disjointed discussion, which is not helped by the different typing and technical abilities mentioned above or by the fact that different respondents may be using different hardware and software that can affect their view of the discussion. We have found that a test run before the actual group can help solve most of these technical problems but problems with typing are harder to solve. All of this means that moderating live on-line groups is very difficult and is at least a two-person job.

Our main recommendation, however, for successful Internet enabled qualitative research is to ensure that all respondents have a true interest in the discussion. We have found that if the topic is of real interest, then the recruitment is easy and the respondents are more thoughtful and detailed than their off-line counterparts.

2. Examples of Future Foundation experience of on-line qualitative research

Example of a moderated on-line group – one-off group with on-line shoppers - 1999

Aim – to look at attitudes to remote purchasing methods held by people who had used our client's on-line shopping service. The client was especially keen to see the on-line methodology demonstrated.

Methodology – respondents were recruited from our client's list of on-line customers. A letter was sent and then follow-up phone calls and e-mails – 10 people recruited in total – expected turnout 6-8.

In practice – Only 4 people turned up to the discussion site at the allotted time. We received 2 e-mails from respondents who said that work commitments meant that they could not attend. Other 'no shows' said on the phone later that other commitments meant they could not attend. One of the four who did turn up, dropped out after 10 minutes because her baby needed feeding and one other left because her machine was too slow and she couldn't keep up with the discussion.

Key lessons learnt – this was a dismal failure! Particularly since our clients were watching. The lessons we learnt were that first, it was foolish to use this method with a group of people who are incredibly time constrained – it would have been much better to run an e-mail group over the course of a week so that they could contribute when they had time. Also, despite using the Internet to shop, the level of interest among this group in discussing the pros and cons of shopping via different methods was not very high.

Finally, while we did ask these respondents to visit the chat room beforehand to check that everything was working, this disaster made us realise that the onus has to be on us to organise a test run/training. In this way we can check that all respondents are up to speed before the group takes place.

Example of a moderated on-line group – week-long group - 1998

Aim – to examine responses to specific web sites and to examine people's ability to complete specific tasks. Again, our clients were enthusiastic about the methodology.

Methodology – respondents were recruited from the Pete Comley's (Virtual Surveys) panel of Internet users. Each respondent was asked to complete a set task prior to the first on-line group – the tasks involved visiting web sites and finding information. The task and web site were discussed in the first group. At the end of the first group a second task was set in a different web site. The group reconvened everyday over the course of a week – five discussions each of 45mins took place overall and 5 different web sites were examined.

In practice – A real success story. Fourteen respondents were recruited, eight of whom managed to come to every discussion. A great group dynamic emerged during the course of the week. By the end respondents were swapping e-mail addresses and personal stories. There was a high level of interest among the respondents. In their eyes they were all early adopters of an exciting new technology (in early 1998 only 10% of the UK population were online) and to them the topic was interesting in its own right. They were all technophiles and we had no problem with software or hardware during the week.

Key lessons learnt – if respondents are interested and the group dynamic is right, they will be willing to take part in a time-consuming project that could not be attempted through other means.

Example of a moderated e-mail group – 2000

Aim – to track a week in the life of Internet users, in terms of what they do on the Internet and also look at their use of other channels. Yet again our clients were very keen to use this methodology.

Methodology – Eight respondents were recruited using traditional off-line methods. Their e-mail addresses were collected and an initial test e-mail was sent out the week before the discussion took place. Initial introductory details were collected, summarised and then sent back to each respondent with a further couple of questions. During the following week, the same procedure took place everyday with the answers being summarised and sent out again the following day.

In practice – This just proved too big a commitment. At the beginning of the week the respondents answered in detail but towards the end of the week we got mainly short, brief responses.

Key lessons learnt – There was no group dynamic in this case to spur the respondents on. Despite having a summary of each other's thoughts to read everyday, there was no interaction between respondents as there was in the case of the weeklong-moderated chat room. Further, the topic was not interesting enough to keep interest going for a whole week. We would have achieved a better result with a 3-day e-mail group covering more ground on each day.

About the Author

Charlotte Cornish

Charlotte manages the Future Foundation's ad hoc and continuous research. Her area of special interest has been looking at the impact of new media on consumers' lives and in making effective use of new research techniques to allow us to assess this change.

Charlotte is an experienced researcher and consultant who has worked at research agencies BMRB and RSL, the Henley Centre and BT. During her time at the Henley Centre she managed its value-added research function, designing and buying research from third party suppliers for its consultancy and consortia projects. She was also project manager on the Centre's flagship Planning for Social Change programme and she also worked on numerous consultancy projects.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

Designing Lengthy Internet Questionnaires: Suggestions and Solutions

Tamara Wojtowicz

Abstract

Good research practice suggests that questionnaire length should be kept to a minimum in Internet research. Jeavons (1999) carried out research on three on-line surveys: the longest of these had the highest percentage of failed sessions. But how can we keep response rates high when the questionnaire is necessarily very lengthy?

In this paper we relate our experiences in maximising response rates for very long (over 100 questions) questionnaires and consider the ways in which these methods may be improved in the future. Many of our projects, especially in the Human Resources and Healthcare fields, have consisted of questionnaires of more than 50 questions in length. By using what we have learnt from analogue reports and from response rates, we have introduced a number of ways to facilitate survey completion for these long questionnaires and increase response rates. In particular:

Presentation

The layout and presentation style chosen for particular groups of questions can often make question blocks appear simple or complex to answer.

Routing

Creative routing can eliminate many unnecessary questions, making questionnaire completion easier and quicker.

Text substitution

If questionnaires ask the same set of questions for a number of different phenomena e.g. particular types of medication, or a number of different job descriptions using text substitution in subsequent pages can shorten the questionnaire and, additionally, facilitate analysis for the researcher.

Saving and loading question responses

This allows a respondent to complete as much of the questionnaire as he/she wishes to in one session. The respondent can then return to the same PC to continue entering data into the same questionnaire at the page that they left. If the respondent needs to search out the information while completing the questionnaire this is an essential function.

> *Identifying structures*
>
> If the respondent must answer several questionnaires where some sections may be near-identical, once the first set of responses are submitted, selected sections may be pre-filled with previously entered replies.
>
> *Menu bar linking to pages*
>
> A menu bar can be inserted showing links to page numbers, or question ranges on each page. A menu can be used along with the above.

This paper gives examples of these methods and provides guidelines for best practice for designing long questionnaires on the Internet. It will be of interest to all those considering the use of Internet and intranet hosted forms for collecting record-based and other informational questionnaire data. It will also be useful for those for whom the advice 'keep it short' is simply not an option.

1. Introduction

Good Internet research practice suggests that questionnaire length should be kept to a minimum. Its users see the Internet as a fast if not instant medium no matter what they are utilizing it for. A one page static questionnaire is seen by many as fast and easy to download (Burckhardt & Guillanton (1998), Page & Wojtowicz (2000)). Nevertheless, some survey projects demand a lengthy questionnaire. The Internet can be the ideal research medium because of its convenience (respondents can complete questionnaires at any time) and geographical reach.

> *"Web interviewing to collect data makes perfect sense in many contexts, such as employee surveys; business to business (especially IT-related fields); among schools and university populations; or other groups where Internet access is readily available or users can be reached on the organisation's own intranet."*
>
> *(Macer, 2001)*

An on-line questionnaire demanding 30 minutes of an individual's time is quite unusual and may not be responded to immediately. The longer the questionnaire and the greater the number of screens the more likely the respondent will terminate the web interview (MacElroy, 2000). However, the technology available (particularly the use of JavaScript programming) can mean that these questions can be asked more efficiently and overall questionnaire length can be reduced. The following techniques should make the completion of lengthy questionnaires more straightforward and less time-consuming They are focused on increasing response rates, providing the respondent is sufficiently incentivised to fill out the questionnaire in the first place. Where examples have been used we have anonymised our clients' identities for confidentiality purposes.

2. Presentation

To a certain extent the design of an on-line questionnaire is similar to the design for other methodologies: paper-based mail survey for data entry or scanning or CATI. The thought process behind the questionnaire is the same and the layout and presentation should be simple and easy to answer for all methodologies. If the questionnaire is not of interest to the individual completing it, he/she will terminate the interview. Clear instructions that give the respondent some idea of how long the questionnaire will take to complete are also important.

However, with traditional mail based surveys the respondents will all receive the questionnaire presented on the same medium: one or more identical paper pages. Respondents to an on-line questionnaire will have different monitor screen size, screen resolution, font sizes, and different Internet browsers.

There are two ways that the survey designer can ensure that the questionnaire is simple and easy to answer for every respondent:

Make the questionnaire easy to view and complete for everyone.

The following should be avoid when designing the questionnaire:

Grid type questions with a large number of code answers

Survey respondents may have to scroll across to see answer boxes at the right hand side if their screens are small or if they are viewing the questionnaire in a minimised window. If the respondent does not notice the scroll bar they may miss one or more options on the grid. In the worst-case scenario some respondents may not be able to scroll across the questionnaire if the questionnaire is a fixed size. In other words, grids are difficult to complete on-line (Dillman, 2000, Jeavons, 1999).

Arranging questions side by side

When the researcher has many questions to ask there may be a temptation to lay the questionnaire out in two columns to “save on space”. This is perfectly acceptable in paper-based questionnaires but as with grid type questions certain respondents may miss the questions in the right hand side of the questionnaire altogether.

Putting the “Next” button on the questionnaire frame

If the “Next” button appears on the frame of the questionnaire, the respondent could click on it to move to the next survey page without seeing all the questions on the page. The “Next” button should appear underneath the questions on that page to encourage respondents to read all the questions. This would not apply if there were only one question per page but with long questionnaires it is likely that a number of questions will appear on each page. Alternatively, the respondent could be forced to answer the questions before proceeding with a dialogue box that indicates that they have not answered certain questions when they click on the page. These “forced answers” should be kept to a minimum in long questionnaires since a dialogue box that appears on every page will increasingly annoy the respondent.

Including unnecessary graphics

Graphics should only be used then appropriate since they increase download time. Download time with long questionnaires can take some time, depending on the speed of the respondent’s modem. This does not stop the survey designer from adding colour since this is only code and consequently small in size. The colour palette of 216 “web-safe” colours should be employed since these colours are less likely to distort on different monitor screens.

Test the survey on different screen sizes, resolutions and font sizes on a number of different browsers.

Netscape Navigator and Internet Explorer are the two most common browsers, used by over 95% of Internet users worldwide. If the survey is designed for a particular client for their Intranet e.g. in the case of a staff survey, the testing procedure will be easier. It is likely that the IT department in the client company will know the PC and browser specifications of the target respondents.

3. Specialisation

Routing

The survey is broken into sequential pieces. These pieces allow the control of program flow on pages such as routing or skip patterns. The logic capability to drive routing can be applied within the software before publishing the questionnaire for on-line publication.

Routing can shorten a lengthy questionnaire by only asking the questions that are relevant to the individual completing the questionnaire. If asking a dichotomous question where the options are "Member" or "Non-member", the next page will show a section of questions that are only relevant to that subset of respondents.

Text substitution

If the same questionnaire must be asked several times for different phenomena e.g. job titles or particular types of medication, then using text substitution in subsequent pages can make questionnaire set-up, completion and analysis easier.

Rather than creating a number of questionnaires for each phenomenon, a single questionnaire can be set up. This reduces the time and consequential cost to the client as well.

For example, say a client wishes to ask about various aspects of a number of different job roles. The respondent is asked to enter what the job title is that they are completing that particular survey for. This literal or open-ended response is then inserted into the following page or pages either as part of the question or as a title to remind the manager which job role details they are completing. When the responses are returned to the researcher, analysis is simplified since setting up a derived variable to look at the different job titles can use data from a single questionnaire.

Selecting and deselecting lists of possible answers

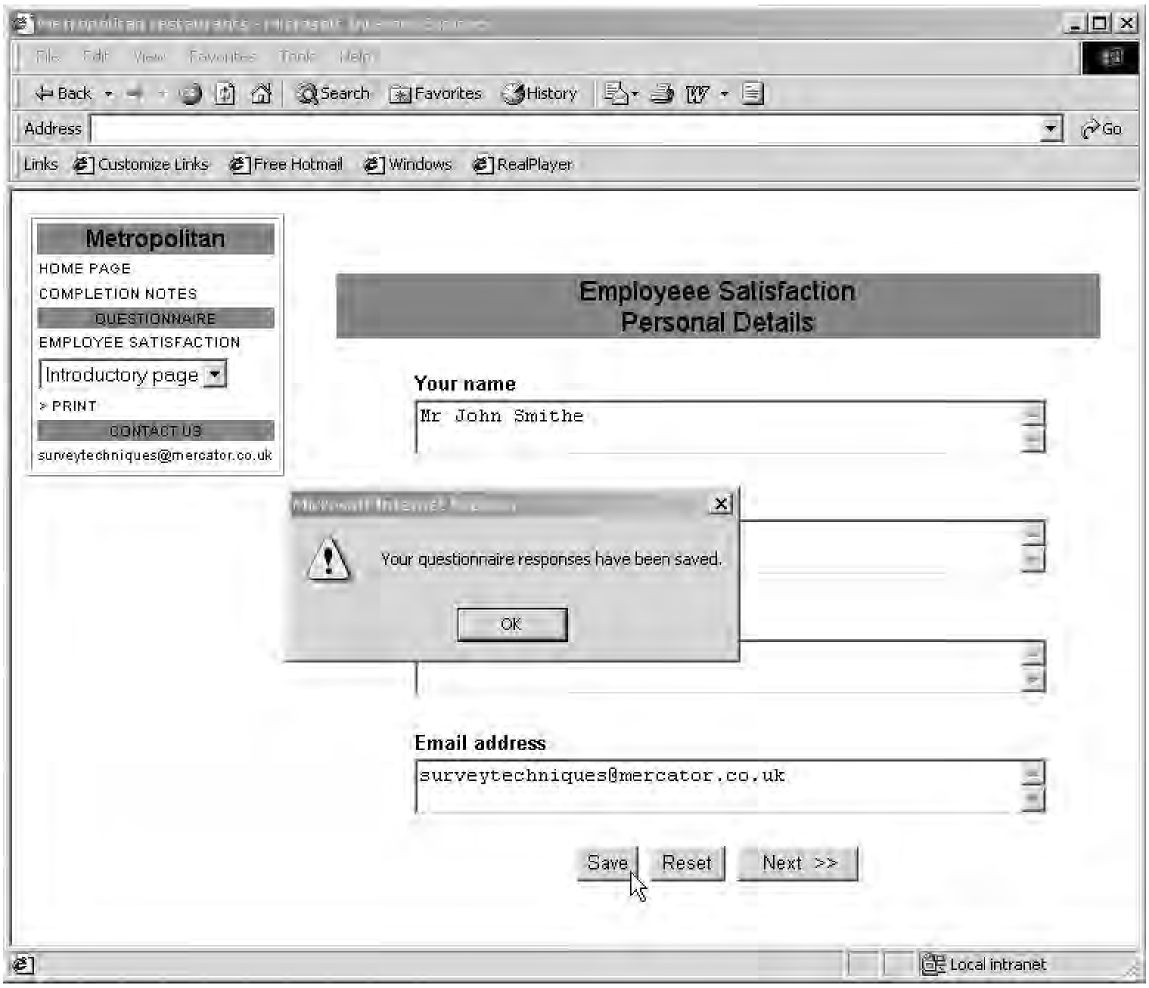
If a respondent selects a number of possible answers, in later pages of the questionnaire certain answers will not appear based on these earlier choices. For example, we know that the respondent has a managerial role in their organisation by their answer to a question on job grade. By our prior knowledge of activities that a manager would or would not carry out, we can filter out the questions relating to inappropriate activities e.g. secretarial activities would not appear.

4. Saving and loading question responses

Completing the questionnaire in more than one session

Lengthy questionnaires cannot always be completed in one session. The busy chief executive or General Practitioner may be interrupted or information must be found that is not immediately to hand. Such interruptions in completing the survey call for a method of leaving the questionnaire and returning to it without losing the data that has already been entered.

We have added JavaScript programming to several lengthy questionnaires to allow this functionality. At the bottom of each "page" of questions a "Save" button is added. Clicking on this button saves the responses already entered. When the respondent returns to his or her PC after their meeting, or finding the necessary data, they type in the address of the survey location. The survey pages will load up and a dialogue box will ask the respondent "You have data already saved for this survey, do you wish to use it?". If the respondent clicks "OK" he/she will be taken to the page where they left off and the responses that they saved earlier will load into the questionnaire. If the respondent clicks on "Cancel" he/she will be presented with a fresh questionnaire containing none of the earlier saved data.


File Edit View Favorites Tools Help
Back Search Favorites History
Address
Go
Links Customize Links Free Hotmail Windows RealPlayer
Metropolitan
HOME PAGE
COMPLETION NOTES
QUESTIONNAIRE
EMPLOYEE SATISFACTION
Introductory page
> PRINT
CONTACT US
surveytechniques@mercator.co.uk
OK
Save
Reset
Next >>
Local intranet

Employeee Satisfaction
Personal Details

Your name

Mr John Smithe

Your questionnaire responses have been saved.

Email address

surveytechniques@mercator.co.uk

This method uses cookies to store the answers on the respondent's PC. When the respondent clicks on "Save" a cookie is added to a cache on their hard drive. There is some debate about the ethical use of cookies and whether they are intrusive for users. However, this issue is avoided since clear instructions tell the respondent that the cookie will be used and he/she chooses whether or not to use the save and load function. Saved data can alternatively be stored on the client's server with a key recorded in a cookie. The key presented to the user on "save" and requested for any subsequent "load" operations.

Identifying structures

If the respondent must answer several questionnaires where answers for certain sections are identical, once the first set of responses are submitted, these sections can be pre-filled with the earlier responses. This method is an extension of the save and load function described above and also utilises cookies for the functionality. The researcher must be sufficiently aware of the subject area being surveyed to pick out sections where responses are common between questionnaires.

Having previously submitted questionnaire responses, when the respondent types in the survey address a dialogue box appears once the questionnaire pages are uploaded. The dialogue box tells the respondent that he/she has responses saved for certain sections of the questionnaire and asks if they wish to use them. If the respondent clicks "OK" the previous responses already appear in those sections of the questionnaire and the respondent can click "Next" without having to complete the questions. However, if there are a few differences the respondent can make those changes too.

Calculating responses where possible

An example of this could be when asking an employee what percentage of time he/she spends carrying out a variety of tasks. If their responses add up to more than 100%, the program could first check the total, and then give the respondent an option to factor the response to reduce them proportionately. The danger of this is that respondents may choose this as an easy option rather than going back to review their answers.

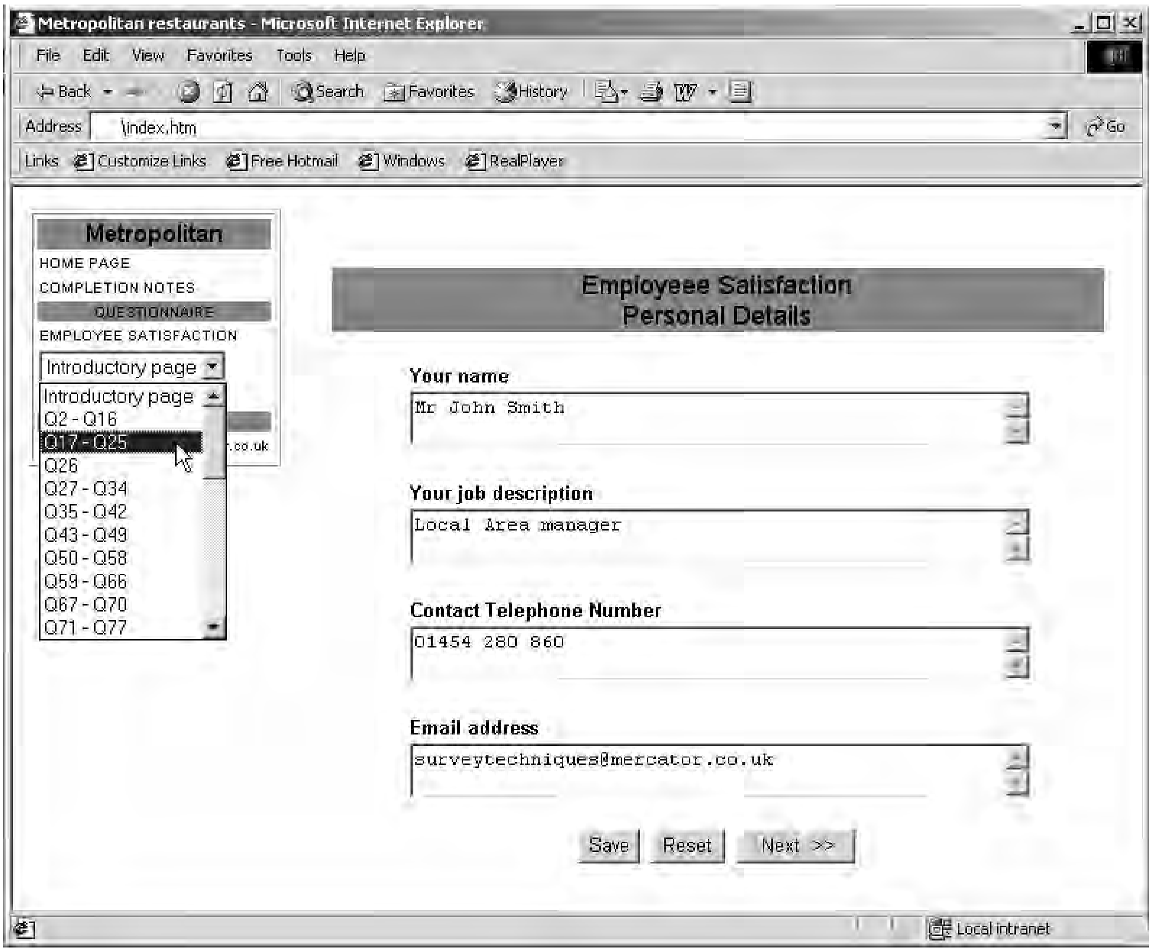
5. Simplifying navigation and providing feedback

Removing the browser toolbar

This can aid the completion of the questionnaire. The browser toolbar is removed and the only controls are those related to the survey completion. This technique is more commonly seen in pop-up surveys.

Menu bar linking to pages

The software that we use produces a layer of pages within the three HTML pages that actually make up the dynamic survey. Consequently, pressing the "Back" and "Forward" buttons on the browser tool bar will not move the respondent from page to page. To facilitate navigation, we can insert a small menu bar at the left-hand side of the survey pages. This menu can either have page or question numbers and respondents can fill out the questionnaire in the order that they wish. This is most useful with lengthy factual questionnaires where order effect is not an issue.



Employeee Satisfaction
Personal Details

Your name
Mr John Smith

Your job description
Local Area manager

Contact Telephone Number
01454 280 860

Email address
surveytechniques@mercator.co.uk

Progress bar

Typically, progress bars are usually shown as a proportion of pages already presented (p) to the total number of pages in the questionnaire (P). In other words, the result of the formula: p/P.

In large questionnaires, which by definition have a large number of questions and thus will tend to have larger chunks of questions omitted, the formula p/P could give some widely inaccurate results particularly towards the end of the completion process. At this time there is also an increased likelihood of respondent dropout when he/she compares the time and effort already spent on the questionnaire to the progress shown.

If we change the definition of P to be the total number of pages that might possibly be asked of any particular respondent, then P will tend to reduce as p increases regardless of the routing employed.

For example, if we had a questionnaire consisting of 50 pages but due to a routing pattern certain respondents only completed pages 1, 2-25 and 50 while others only completed pages 1, 26-49 and 50. This ordering is not unlikely since one of the aspects of good questionnaire design is to keep related questions together on the questionnaire. When a respondent from the first group reaches page 25 he/she will see that the progress is only 50% whereas they are actually on page 25 out of 26. Most people expect the progress bar to proceed in a linear fashion. A side effect of this is that the first group of respondents is more likely to give up completing the survey and therefore will be underrepresented in the final results.

6. Summary

In this paper, the author has enumerated a number of techniques and guidelines for the survey researcher wishing to publish a long questionnaire on the Internet. These techniques have many applications for a wide range of surveys and should enhance current research requiring lengthy questionnaires.

Future directions

Technology is advancing at an incredible rate. Sophisticated applications are more and more commonplace on the Internet. As the technical know-how of survey respondents increases we will be able to more efficiently employ the techniques above. But there is still much to be done in developing and utilising further web survey techniques. The Internet researcher must keep up to date with latest changes in technology as well as any ethical issues related to them. Although, technology is moving along at a rapid pace, the on-line survey still has as it's interface a human respondent. We are still using these tools to achieve the same goal - to efficiently capture a respondent's answers.

References

Burckhardt, J & Guillanton, P. (1998) ***Surfing the Learning Curve: Examining the Applications of a Major Pan-European Internet Survey for Yahoo!*** ESOMAR Seminar Proceedings on the Internet and Market Research, ESOMAR Amsterdam, The Netherlands.

Dillman, D. (2000). ***Mail and Internet Surveys: The Tailored Design Method.*** 2nd Edition. John Wiley & Sons Ltd.

Jeavons, A. (1999). ***Ethology and the Web: Observing Respondent Behaviour in Web Surveys.*** ESOMAR Worldwide Internet Conference, Net Effects, ESOMAR Amsterdam, The Netherlands.

MacElroy, B. (2000) ***Variables influencing dropout rates in Web-based surveys.*** Article Number: 0605 Quirk's Marketing Research Review (www.quirks.com).

Macer, T. (2001) ***Get connected*** Research Guide to Internet Technology. A supplement of *Research* Magazine. London, UK.

Page, B. & Wojtowicz, T (2000) ***Benchmarking on a Global Scale: A Case Study Illustrating Experiences and Implications of the Use of the Internet.*** ESOMAR Seminar Proceedings Net Effects 3, ESOMAR Amsterdam The Netherlands.

About the Author

Tamara Wojtowicz

Research Consultant

Mercator, 5 Mead Court, Thornbury, Bristol. BS35 3UW. ENGLAND

Sampling and Instrumentation Issues

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

The Promises and Perils of Web Surveys[1]

Mick P. Couper

Abstract

This paper presents an overview of key issues related to Web surveys. Various types of Web surveys will be classified and compared within a total survey error framework, focusing on sampling, coverage, non-response and measurement error issues. The goal of the review is to separate the rhetoric both for and against Web surveys from the objective criteria that can be used to evaluate the efficacy of a particular approach relative to alternative methods of surveys data collection. Empirical evidence will be brought to bear to better understand the potential uses and possible dangers of Web surveys. Web surveys are indeed an exciting new weapon in the survey researcher's arsenal, but there is much we need to learn about how to make effective use of this method.

Keywords

Web surveys; coverage, non-response; measurement error; design

1. Introduction

The Web is changing the face of the survey industry like few other innovations before it. Much of this could be attributed to the overall speed of development of the Web in general. Yet, for all the excitement and attention being paid to Web surveys, much of the rhetoric surrounding the introduction of the new method has a familiar ring to it. In the same way that telephone surveys in general—and CATI in particular—were hailed as the new faster, cheaper, better way of conducting surveys, and similar arguments were made for CAPI (see Couper and Nicholls, 1998), Web surveys are being promoted as the survey technology of the new millennium.

One of the more outspoken proponents of Web surveys is Gordon Black, CEO of Harris Interactive. In responding to critics of the online poll, Black stated "It's a funny thing about scientific revolutions. People who are defenders of the old paradigm generally don't change. They are just replaced by people who embrace new ideas." (Wall Street Journal, April 13, 1999; see also Mitofsky, 1999). Black is not alone in claiming that Web surveys will supersede other modes of survey data collection.

[1] Alternatively titled, "The Good, The Bad, And The Ugly." Parts of this paper are based on an earlier review published in *Public Opinion Quarterly.*

With the heady rhetoric and hyperbole surrounding the advent of this new method, it is important to reflect on the claims and counter-claims in the context of familiar evaluation criteria and relative to tried and tested methods of survey data collection. Web surveys are neither the answer to all survey researchers' prayers nor are they the dire threat to the future of the survey industry some fear. The truth, as always, lies somewhere in between.

Above all, when talking about Web surveys, what is needed is perspective. Like the parable of the blind men and the elephant[2], Web surveys are viewed in different ways by different people. The Web is often variously represented either to emphasize its enormous potential or to draw attention to its not inconsiderable drawbacks as a tool for survey research.

This paper offers a brief overview of some of the benefits and pitfalls of Web surveys as they are currently being implemented. I review some of the findings in the rapidly expanding research literature on Web surveys. The key premise of the paper is that Web surveys, as with any other method of survey data collection, must be evaluated in context. The framework I use is the total survey error perspective (see, e.g., Groves, 1989), a very useful tool for organizing such an exercise. I first offer some general observations on Web surveys and data quality, then review what we currently know about Web surveys in terms of each of the key sources of error in turn.

2. Web Surveys and Data Quality

To elaborate on the importance of context in evaluating or discussing Web surveys, three points can be made about any comparisons of Web surveys to other methods of survey data collection:

1. The standards of comparison should be made explicit. For example, if the comparison is to a mall intercept survey, Web surveys do indeed offer the opportunity to collect much more (and richer) data at far less cost. On the other hand, if the comparison is to a high response rate probability sample of the general population, the failures of the Web method become rapidly evident. The question here is, To what are we comparing the Web survey?

2. The criteria for evaluation should be made explicit. Web surveys, like all other methods of survey data collection, involve trade-offs. Survey methodology has a well-developed model of total survey error that can serve as a useful departure point for such comparisons. In terms of costs, Web surveys may be unequaled. But such benefits often come at the expense of low response rates, poor coverage of the general population, and the difficulties of selecting probability samples. So, here the question is, What source of error is the focus of attention?

3. The type of Web survey should be made explicit. As I have discussed elsewhere (Couper, 2001), there are many different ways to implement Web surveys, and some are better (by some of the evaluation criteria) than others. For instance, probability samples (albeit of limited scope) *can* be developed for Web surveys, but these may suffer from other problems such as high nonresponse. Here we should ask, What type of Web survey are we talking about?

[2] Six blind men encounter an elephant and each, feeling different parts of the elephant, reaches different conclusions about what an elephant is. See the poem by John Godfrey Saxe (1816-1887).

The point is that any discussion of the blessings or evils of Web surveys requires explicit details of both the source and target of the comparison, and of the criteria being used to compare the two. The way one asks the questions about Web surveys may determine the likely answer.

O'Muircheartaigh (1997, p. 1) offers as a definition of error in surveys, "work purporting to do what it does not do." He goes on to write, "Broadly speaking, every survey operation has an objective, an outcome, and a description of that outcome. Errors (quality failures) will be found in the mismatches among these elements." Survey quality is not an absolute, but should be evaluated relative to other features of the design (such as accuracy, cost, timeliness, etc.) and relative to the stated goals of the survey. The relative quality of a particular approach must be judged in light of alternative designs aimed at similar goals and with comparable resources. For example, comparing Web surveys to quota samples, mall intercepts, low response-rate RDD surveys, magazine insert surveys, customer satisfaction cards, and so on, is likely to lead to different conclusions than comparisons to high-response rate interviewer-administered surveys of the general population.

More so than any other mode of survey data collection, the Internet has led to a large number of different data collection uses, varying widely on several dimensions of survey quality. Each Web survey must thus be judged on its own merits. Space does not permit a detailed discussion of different types of Web surveys here (these are discussed at greater length in Couper, 2001). However, I offer a brief summary of the key distinctions among various types of Web surveys.

A fundamental distinction can be made between non-probability and probability-based methods. The former rely on self-selected or volunteer participants, and members of the target population do not have known, non-zero probabilities of selection. Hence, inference or generalizations to that population are based on leaps of faith rather than on established statistical principles. On the other hand, probability samples are a necessary but not sufficient condition for inference to a larger population. Coverage, non-response and measurement errors may threaten the representativeness of these designs.

Non-probability approaches dominate the Web survey world. These range from "question of the day" instant polls that make no claims of scientific validity to pre-recruited or opt-in panels claiming to be representative of or projectable to the general population. While surveys of the first type are relatively harmless, they may be mistaken for legitimate surveys and hence cause confusion. A number of Web surveys recruit respondents through banner ads or other open invitations, often with no controls on multiple completions or "ballot-stuffing." Unlike the first type, these latter surveys are often being done by academic organizations and others with the aim of making generalizations to some larger population. Two prominent examples are the Georgia Tech University's series of GVU surveys (see Kehoe and Pitkow, 1996) and the National Geographic Society's Survey 2000 (see Witte, Amoroso and Howard, 2000).

The creation of opt-in panels for Web surveys is a fast-growing sector of the (Web) survey industry. There are literally dozens of survey organizations around the world that are actively soliciting members for their panels[3]. It is these panels that are generating the most interest (and criticism) from traditional survey researchers, in part in reaction to the claims being made by the developers of such panels.

[3] There is even a Web site devoted to listing these sites: http://www.money4surveys.com

Arguably the best known of these in the U.S. is Harris Interactive's Harris Poll Online, which claims over 7 million members in its panel. Harris Interactive's Web site notes, "... we are able to survey much larger numbers of consumers than could be done cost-effectively using telephone or mail techniques, *producing results that are much more reliable and projectable*" (emphasis added). Greenfield Online makes claims of being "the world's largest Internet_based marketing research panel," with over 500,000 registered respondents representing households containing over 1.7 million individuals. NFO Interactive's Website had a similar boast: "We at NFO Interactive are leading the market research industry into this brave new world.... Today, we have the largest, representative interactive panel in the world," which is claimed to be a "*fully representative panel* of nearly 260,000 (and growing) interactive households and over 750,000 interactive consumers" (emphasis added).

As indicated by these statements, these companies are clearly trying to position themselves as leaders in this burgeoning field. Although the claims can be viewed as marketing hyperbole, the danger lies is the potential inability of the average client of such research to understand what "representative" or "projectable" means, and to evaluate the impact of these approaches on the research they commission. Reacting to the growth of opt-in Web survey panels, Warren Mitofsky noted: "My view is that it's an enormous threat....There are too many unsophisticated people willing to pay, and this kind of bad research is going to drive out some good research." (Associated Press, May 19, 2000)

Despite the dominant position that such non-probability surveys have appeared to gain in the Web survey world, it is not only possible to do probability-based Web surveys, but many are indeed developing, evaluating and implementing a variety of different approaches to do just that. Indeed, Knowledge Networks (formerly InterSurvey) in the U.S. and (on a smaller scale) CentERdata in the Netherlands are demonstrating that it is even possible to develop and deploy probability-based methods for Web surveys of the general population.

The key challenges for developing probability-based Web surveys are coverage of the population of interest and the development of a sampling frame. I return to these issue below. The problem is being solved in several ways: (1) survey only the Internet population (however that may be defined), using some other method (e.g., RDD) to identify and recruit sample persons; (2) provide Internet access to the general population in exchange for participation in Web surveys; (3) restrict studies to known populations with high rates of access, (4) conduct intercept- or transaction-based surveys of Internet or Web users; (5) use Web surveys as one option in a mixed-mode design. Each of these approaches has particular attractions for certain types of research, but all also have potential drawbacks, most notably related to non-response.

Further details of the different approaches are discussed elsewhere (see Couper, 2001). I now turn to a discussion of various sources of error in Web surveys, focusing primarily on coverage and sampling error, non-response error, and measurement error.

3. Coverage and Sampling Error

Probability sampling assumes that each person in the target population has a known, non-zero probability of selection. The *known* selection implies that a sampling frame can be developed. This could be a list of all members of the target population of interest, or a procedure for identifying and selecting sample elements, as in transaction-based approaches.

A sampling frame of all Internet users does not currently exist, and in unlikely to do so in the foreseeable future. While lists for some subset of users (e.g., subscriber lists for Internet service providers) may be obtainable, the format of e-mail addresses (if we argue this is a suitable proxy for Web use) prevents the development of an RDD-like procedure for sample selection. The sampling frame problem is similar to the reason mail surveys of the general population are typically not undertaken in the U.S. However, for more restricted populations, such lists may be much more easy to come by, making probability sampling feasible. These include members of professional associations, student populations, and so on.

There are essentially three approaches to sampling in Web surveys. These can be summarized as follows:

1. No sampling. Many Web surveys use open invitations on portals or Websites, with no control over who completes the survey or how often they do so. In these cases there is no way to determine the denominator for calculating a sampling fraction or a response rate.
2. List-based sampling. The sampling process is relatively easy. The coverage depends on the quality of the list. For opt-in panels, the sample for a particular survey can be representative of members of the list, but not of the general population. The list typically consists of e-mail addresses as the means of invitation/delivery, although some surveys have used mailed invitations to a postal address to solicit participation in a Web survey.
3. Intercept methods. These include the use of banner advertisements and pop-up invitations. In each case the population is defined in terms of visits (or visitors) to the site, and sampling fractions can be the number of exposures to the ad or visitors receiving the invitation.

Generally the sample selection issue itself has been of little concern in Web surveys. If a list exists, the procedures for sampling from such a list and estimating the probabilities of selection are well known. Sampling for intercept-based approaches is also relatively straightforward. Where no list exists, either other methods are used (e.g., RDD) or probability sampling is abandoned altogether.

It is the second attribute of probability methods—that members of the target population have a *non-zero* chance of selection—that has received the most attention with respect to Web surveys. This is typically referred to as a problem of coverage. Coverage error occurs to the extent that members of the target population have no chance of being selected in the sample, whether because they have no access to the Web or because of errors in the list or sampling procedures. Coverage error is a function of both the proportion of the target population that is not covered by the frame and of the difference on the survey statistic between those covered and those not covered. This can be represented as follows:

$$Y_c = Y_t + \frac{N_{nc}}{N_t}(Y_c - Y_{nc})$$

Where Y_c is a statistic estimated on the covered portion of the population, Y_t is the statistic we wish to estimate for the full population, N_{nc}/N_t is the noncoverage rate, and Y_c - Y_{nc} the difference between the covered and the not covered on the statistic of interest.

Many are focusing solely on the first part of expression (the coverage rate), while others are looking at the second part (the difference term) in terms of demographic variables only. Neither response adequately addresses the coverage problem. Coverage error can occur if *either* part is different from zero on the variables of interest.

Several themes are evident in the arguments being used to counter concerns of coverage. These can be briefly summarized as follows:

1. Other survey methods (especially telephone surveys) are not projectable because of low response rates or coverage problems, so the Web is no worse. For example, Ahlhauser (1999) writes, "From traditional research methodologies we know that no sampling is truly random. With declining participation rates in all the intercept methods, their projectability is subject to renewed scrutiny." Similar arguments are made by the owner of the SURVEY.NET polling site: "Traditional polls rely on a 'human factor' and a supposed 'random sample' from which to generate results. It has been argued that traditional polls could be more accurate than the SURVEY.NET system... I challenge anyone to show me any survey which endeavors to collect accurate information which isn't biased" (http://www.survey.net/sv-faq.htm). Similarly, Terhanian and Black (1999) give as their *raison d'etre* for moving to the Web the high refusal rates for telephone surveys[4].

2. Web penetration is increasing every day, and the Web population increasingly resembles the broader population demographically, so the coverage problem is a temporary one that will soon disappear. Several proponents of Web surveys note that the same concerns were raised with the introduction of telephone surveys, but these concerns have largely disappeared. Again, Ahlhauser (1999) notes: "Concern about the projectability of Web research follows the grand tradition of concern for telephone interviewing when it emerged. Telephone interviewing overcame these concerns by dint of the incredible advantages it brings to many types of surveys, and propelled by the pervasive and transforming effect the telephone was having on society." Black and Terhanian of Harris Interactive have made similar arguments. Most of the focus of these arguments is on age and gender distributions, not education, income, race or key substantive variables of interest. For every claim that the Web population resembles the general population, there are studies showing the persistence of the "digital divide."

3. Weighting can solve any problems that might exist. For example, Cohen (1999) stated in response to those who argue that Web surveys are not projectable, "Samples can be weighted to conform to given populations." Terhanian (1999) stated the following: "In principle, we at Harris Interactive believe that there should be no difference aside from sampling error between survey response elicited through Harris Poll telephone research and Harris Poll Online Research as long as both surveys: ask the same questions; occur at the same time, and draw samples at random from the exact same population *or are thoughtfully and appropriately weighted*" (emphasis in original).

4. It's the result that counts, and the obtained from Web surveys resemble those from other survey methods. For example, Yoffie (1998) writes: "When our clients have asked us to run similar parallel studies, the results have been virtually identical. Others in the field have reported similar findings." Similar claims are made by Terhanian (2000, p. 8): "...we have conducted more than two hundred surveys with parallel telephone and Internet components in the past two years. The results of these surveys are indistinguishable in most subject areas."

[4] It's interesting that, given these arguments, Harris Interactive uses the very same telephone surveys they dismiss as the benchmark for adjustment of their Web surveys.

5. It's size that counts. This argument implies that numbers are more important than quality, or alternatively that large numbers alone are an indicator of quality (i.e., reliability). For example, Kehoe and Pitkow (1996) write about the GVU surveys: "Since we use non-random sampling and do not explicitly choose a sample, having a large sample size makes it less likely that we are systematically excluding large segments of the population. Oversampling is a fairly inexpensive way to add more credibility to a non-random Web-based survey." The following statement appeared on the National Geographic Survey's website until protests by survey researchers led to its removal: "We received more than 50,000 responses—twice the minimum required for scientific validity..."

Turning again to the issue of coverage of the general population, the Web survey industry (and indeed the broader "dot.com" sector) has a vested interest in inflating the estimates of the number of Web users. For example, a recent (February 22, 2001) press release by Nielsen//NetRatings had the following headline: "More than 169 Million People Have Internet Access" in the U.S. In the body of the statement, it says that Internet penetration has reached 60% percent in the U.S. as of January 2001. Given a total U.S. population of 281 million as of the 2000 census, this must mean that every man, woman and child is included in the denominator of this estimate[5]. If we use as a base the population 15 years old and older, the number of Internet users should be closer to 130 million.

The 60% estimate obtains some support from an RDD telephone survey conducted by the Pew Internet in American Life Project in November/December 2000, which produced an estimate of 56% of adults online. But, these are both likely overestimates of the U.S. online population. U.S. Census Bureau data from August 2000 indicate that 44.4% of persons 15 and older in the U.S. use the Internet (NTIA, 2000). This yields an estimated 116.5 million Americans online, a far cry from the 169 million number touted by Nielsen//NetRatings.

Furthermore, with regard to coverage, it is not the *number* of people with Internet access, but the *proportion* of the relevant population, that is important. To illustrate this point, China is currently not far behind the U.K. in terms of numbers of Internet users (16.9 million for China and 19.5 for the U.K., according to Global Reach (http://www.glreach.com), but in terms of proportion of the population, the numbers are very different. With an estimated population age 15 and older of around 950 million for China and just over 19 million for the U.K., this represents only 1.8% of the Chinese population and 40% of the U.K. population estimated to be online. This disparity is not stopping several companies from creating Web survey panels in China.

To turn to the second part of the coverage error formula, the *differences* between the haves and the have-nots are critical in terms of likely coverage error for Web surveys. Efforts to dismiss these concerns are also widely seen in the Web survey industry. When, for example, Black claims that "The Internet looks more like America than it ever has," (Wall Street Journal, April 13, 1999), one wonders what kind of America he has in mind?

This assessment is quite at odds with conclusions drawn from the latest Census Bureau study, conducted in August 2000. The report concludes that while tremendous growth in Internet access has occurred "across all demographic groups, including income and education levels, races, locations, and

[5] In addition, the press release suggests that these data are from a panel of Internet users. It's not clear how one estimates Internet penetration using a panel of Internet users. As of this writing, no response has been received from Nielsen//NetRatings to an e-mail request for clarification of the numbers in the press release.

household types. Nevertheless, some Americans are still connecting at far lower rates than others, creating a *digital divide* (i.e., a difference in rates of access to computers and the Internet) among different demographic groups." (NTIA, 2000, p. 2; emphasis in original).

Similarly, the authors of the Pew Internet Project's Internet Tracking Report (February, 2001) conclude that "The increase in online access by all kinds of Americans highlights the fact that the Internet population looks more and more like the overall population of the United States. However, there are still some notable demographic differences when it comes to access." Some of these notable differences are shown in Table 1.

Again we must remind ourselves that it is not the demographic differences between the online and off-line populations that are of primary concern. Of issue for coverage error are all the things we are trying to measure in our surveys: attributes, attitudes, opinions, behaviors, preferences, intentions, etc. Even if the Web population resembles the non-Web population in term of age and gender (for example), this does not mean that the two groups are identical in terms of the substantive variables of interest. If they were, we would be able to perfectly predict attitudes and behaviours on the basis of a small set of background measures.

While it is clear that non-coverage rates and differences between the covered and non-covered populations do not (yet) permit projections to the general population, there are some populations of great interest to researchers that have near-universal Web or Internet coverage. Student populations are one example. In a survey of University of Michigan undergraduate students conducted in 1999, for only 2.4% of the sample could we find no evidence that they had accessed their e-mail during that semester, and 5.2% during the survey period (Couper, Traugott, and Lamias, 2000). In other words, we had positive evidence that at least 97.6% of the sample had used e-mail. This parallels Radler's (2000) finding that only about 4% of University of Wisconsin-Madison students did not have a preferred e-mail address.

For another example, I reviewed the member information published in the 2000 directory of the American Association for Public Opinion Research (AAPOR). Of the 1,590 names listed, 89.7% provided an e-mail address in the printed directory, compared to 91.4% who provided a telephone number. For related professional groups, such as the Association for Statistical Computing (ASC), these proportions are likely to be at least equally high.

Another group of obvious interest is the online population itself. While such a population is difficult to measure and near impossible to sample with any precision, surveys of this population do not suffer from the same coverage problems as those that make claims beyond this group.

In terms of samples of the general population, the sampling and coverage issue remains a key concern. But we should not focus exclusively on this issue alone. On the one hand, without probability sampling, one simply cannot call on the weight of statistical theory and practice to support statements such as "projectable," "generalizable" or "representative." No amount of statistical adjustment or propensity weighting will alter this basic fact. On the other hand, however, simply because a survey is based on probability sampling methods is no guarantee of its representation. Other sources of error (especially coverage and non-response) may still swamp concerns about sampling error, and may produce a sample that is patently dissimilar from the population it purports to represent. This second observation is often overlooked in the heated debate about the scientific value of Web surveys. I thus now turn to these other potential sources of error.

4. Non-response Error

Even if coverage concerns were reduced by limiting the population of inference, non-response error may still threaten the utility of Web surveys for reliable projections to a larger population. Non-response error arises through the fact that not all people included in the sample are willing or able to complete the survey. As with coverage error, non-response error is a function both of the rate of non-response and of the differences between respondents and non-respondents on the variables of interest (Groves and Couper, 1998). Non-response error can be represented as follows:

$$Y_r = Y_t + \frac{N_{nr}}{N_t}(Y_r - Y_{nr})$$

Where Y_r is a statistic estimated on the respondents, Y_t is the statistic we wish to estimate for the total population, N_{nr}/N_t is the non-response rate, and Y_r - Y_{nr} the difference between respondents and non-respondents on the statistic of interest.

As with coverage, a low non-response rate (high response rate) does not guarantee an absence of non-response error. But the lower the rate, the greater the potential for differences between respondents and non-respondents on the statistic of interest to affect overall estimates derived from the sample.

For surveys where the frame cannot be identified, the problem of non-response is hard to define. For example, if an open invitation is issued on a Web portal to participate in a survey, the denominator of those eligible to participate is typically not known, and therefore the non-response rate is unknowable. For some surveys even the numerator is not clearly known because of possible multiple completions (ballot-stuffing) by the same person. Banner-advertised Web surveys are an exception, where estimates of the number of visitors exposed to the invitation can be obtained (through counts of page hits).

The measurement or evaluation of non-response error is generally only tractable in cases where the frame and the chance of selection are known (in other words, probability-based surveys). For a list-based sample, the number of selected sample persons is easy to determine. For intercept-based approaches, the denominator is the number of visits (not necessarily visitors) receiving the invitation. It is not surprising, therefore, that most of the response rate data we have on Web surveys are from list-based samples. Furthermore, many of these response rates are based on explicit comparisons to alternative modes of data collection. Table 2 presents a summary of some of the studies. The surveys summarized in this table are all from list-based samples, where the denominator is known. In several of the examples, sample cases with invalid e-mail addresses were removed from the denominator.

For opt-in panels, details on response rates less clear. Schmidt (2000) of Greenfield Online reported response rates ranging from 20-60% for the online panel, and Terhanian (2000) reported response rates ranging from 20-25% for members of Harris Interactive’s online panel. Schonlau (2000) used the Harris Interactive panel for a survey of California residents: of the 70,932 invited to do the survey, 14% started the survey and 12% completed it. In a recent study using Survey Sampling Inc.’s (SSI) panel, we obtained a 20.3% response rate with an e-mail invitation and single reminder (Couper, Tourangeau, and Steiger, 2001). An additional 2.9% of the sample started the survey but did not complete it. A later survey with no reminder received an 11.9% response rate. Yoshimura and Ohsumi (1999) administered four different surveys to samples of opt-in panel members of three different companies in Japan. The response rates they obtained ranged from 7.7% for a survey on consumer behaviour to 19.7% for a survey about the Internet.

Comley (2000) summarizes response rates to a number of intercept or pop-up surveys conducted by Virtual Surveys. He reports that the response rates range from 9% to 48%, with an average of 24%. One example was a pop-up survey for Thomas Cook, in which every 10th visitor to the site was invited to complete a short survey. Of the 4,500 visitors exposed to the invitation, 22% completed the survey. MacElroy (2000) similarly reports response rates of 15% to 30% for intercept surveys. McLaughlin (2000) reported average response rates of 15% to CyberDialogue's intercept surveys.

Response rates for banner-advertised surveys are even lower. Tuten, Bosnjak and Bandilla (2000) experimented with different appeals (intrinsic versus extrinsic) on four different Web search engines. They obtained click-through rates ranging from 0.13% (13 responses per thousand exposures) to 0.44%. MacElroy (2000) supports this finding with estimates of response rates around 0.5% for banner-advertised surveys.

This brief review of available information on Web survey response rates suggests that non-response is a key concern for Web surveys, especially when compared to alternative methods such as mail. Given the relative per-unit costs of the two methods, this may be of little concern for those whose interest is a large number of respondents, but are of concern for those hoping to make reliable inference to a larger population. Again, it depends on how one looks at these numbers. By some standards, a 40% response rate to a Web survey of college students or a 20% response to a Web intercept survey are quite respectable. But measured against the yardstick of many academic and government surveys, these rates are insufficient to inspire confidence in the method.

Is this a temporary problem until we learn how to exploit features of the new technology to reduce non-response? Or are Web survey response rates doomed to remain lower than comparable mail surveys, in much the same way that telephone survey response rates typically fail to reach the rates for face-to-face surveys of similar design and content? I suspect the latter scenario is more likely. Several explanations may be offered for this conclusion. One is that tried and tested motivating tools used in mail surveys (e.g., advance letters, personalized signatures, legitimising letterhead, incentives, etc.) cannot be implemented in the same way in Web surveys and functional equivalents are yet to be developed and tested. Finding electronic equivalents of response-stimulating efforts is work that remains to be done. For example, the incentives literature suggests that prepaid incentives work better than promised ones, and cash works better than in-kind payments. But how does one deliver prepaid cash incentives in a Web survey without giving up on the other advantages of electronic invitation delivery?

Even if these response-inducing techniques are developed, technical difficulties in using the Web may discourage some from completing (or even starting) the survey, relative to the ease of completing a paper-and-pencil mail survey. Slow modem speeds, unreliable connections, low-end browsers, etc., may inhibit Web survey completion from home. In some countries (and for some Internet providers), connect-time costs may deter people from doing so. In such cases, there may be real costs of completing a Web survey. When given a choice between Web and paper, with a few exceptions, respondents overwhelmingly appear to choose the latter.

Finally, one can speculate that the culture of the Web may militate against high response rates. There are already well-developed normative strictures on the Web limiting unsolicited invitations (spamming). Furthermore, while the development of telephone screening devices is a relatively recent development, such systems are already well established on the Web. Given the fast-paced multimedia characteristic of the Web, which is controlled by the user and offers almost-infinite (and typically far more interesting) alternatives to completing a survey, the task of getting a potential respondent's attention and keeping it throughout the survey may be one of the more daunting challenges facing Web survey designers.

Again, we must remind ourselves that response rates are only one part of the non-response error equation. Exploring the differences between respondents and non-respondents to Web surveys will help us understand the effect that non-response may be having on our estimates. One advantage of list-based approaches to sampling for Web surveys is that one typically has auxiliary information about members of the list. Such auxiliary data could be useful in examining differences between respondents and non-respondents. Another advantage of list-based samples is they permit follow-up of non-respondents, which may help not only to increase response rates but also help us understand why some people choose not to respond (or are unable to respond) to the survey. By way of contrast, intercept-based methods provide little or no information about the selected sample person, making follow-up impossible and leaving little way to explore non-response effects.

To date, very little work has focused on non-response error in Web surveys. Problems of coverage diminish as one increases the restriction on the population of interest to definable groups for which sampling frames or procedures can be developed. But, as the sampling and coverage problems become less salient, the non-response problem is likely to become relatively more important.

5. Measurement Error

Measurement error simply stated is the deviation of the answers of respondents from their true values on the measure. Measurement errors in self-administered surveys could arise from the respondent (lack of motivation, comprehension problems, deliberate distortion, etc.), from the instrument (poor wording or design, technical flaws, etc.), or from some interaction between the two. It is with respect to the measurement process that Web surveys may offer the greatest potential. In some ways Web surveys are like other methods of survey data collection, but in other ways they are unique. It is these common and unique characteristics that provide enormous opportunities and challenges for instrument design.

First, Web surveys are self-administered. In interviewer-administered surveys, well-trained interviewers can often explain unclear terms to respondents, keep them motivated, reassure them of the confidentiality of their answers, probe incomplete or inadequate responses, and so on. In other words, they serve as intermediaries between the researcher and the respondent. In self-administered surveys there is no such intermediary, and the survey instrument itself serves to convey the researcher's questions and expectations to the respondent. Similarly, the respondents' only means of communication with the researcher is through the medium of the instrument, and their answers and intentions must be taken at face value. This places greater focus on the self-administered survey instrument. The instrument must be easy to understand and to complete, must be designed to keep respondents motivated to provide optimal answers, and must serve to reassure respondents regarding the confidentiality of their responses. On the other hand, there are documented advantages of self-administered methods, particularly for the collection of sensitive information and the reduction of social desirability effects, that are likely to apply for Web surveys too.

Second, Web surveys are computerized. In contrast to mail surveys, which are static instruments, Web surveys can make use of the full power of computer-assisted interviewing (CAI) methods. These include automated branching or skipping, randomisation of questions or response options, tailored fills or question wording, range and edit checks, feedback to respondents, and so on. While not all Web survey designs make use of these features, they are nonetheless potentially powerful tools for interacting with respondents, assisting them in the completion of the task, and motivating them to continue with the survey.

In these two respects, Web surveys are much like computer assisted self-interviewing (CASI) methods. However, Web surveys differ from CASI in that there is no trained interviewer present, and the survey organization does not provide the equipment. This latter point means that respondents are completing the survey on a variety of different hardware platforms and using a variety of software systems, with the potential that the instrument does not look and act in an identical manner for all respondents. This again puts more burden on design.

Finally, Web surveys have the power to extend the visual elements of presentation beyond what is usually feasible in paper surveys. Web survey instruments no longer only (or primarily) consist of verbal features (words and numbers) but can also make use of rich visual features. These visual enhancements include still and moving images, animation, line drawings, pictures, colour, shapes, etc., not to mention true multimedia which includes both sound (aural) and pictures (visual) features. The graphical nature of the Web frees the survey designer from the traditional constraints of paper-based questionnaires (order, font, colour, etc.). It's not that these design features could not be used before, but they were expensive and time-consuming to develop and reproduce in large quantities, and were thus used sparingly. In contrast, embedding a colour photograph or image in—or changing the background colour, design or layout of—a Web survey is a relatively trivial task.

All these elements combine to make the Web a unique medium for the presentation of survey questions and for the elicitation of responses. In fact, the Web permits the extension beyond the traditional survey "question" to include a wide variety of stimulus material. While the Web provides a wonderful opportunity to "think out of the box" and expand the variety of ways information can be presented to respondents, this freedom may come at a price. For example, while images are increasingly being used in Web surveys to enhance the user experience and motivate respondents to continue with the survey, the addition of images may have unintended consequences for the survey questions and the responses being elicited. Even when the image is explicitly designed to supplement the question text, the effect may be different than that desired. Thus, with the increased range of tools comes the heightened possibility of inadvertently introducing measurement error.

There is much we need to learn about how visual elements of a Web survey interact with the verbal elements in affecting respondents' answers. Visual images can be used in many different ways in Web surveys, for example:

- Images irrelevant to the survey task (e.g., banner ads), designed explicitly to draw attention but not necessarily related to the survey content.
- Motivational images designed to keep respondents participating in the survey (i.e., prevent dropouts) but orthogonal to question content.
- Images intended as supplemental cues for the question text (e.g., pictures of political candidates accompanying questions of candidate choice).
- Images as core content of the question (i.e, as the stimulus itself, as in brand recognition, landscape preferences, etc.)

The effect that visual images may have on survey measurement may depend not only on the purpose of the image, but also how it is visually linked to the question text on the screen, and the content of the visual image itself. The psychological evidence is that verbal and visual information are stored and processed in parallel (e.g., see Paivio's, 1986, dual coding model, which argues that both working and long-term memory uses both types of representation). The presence of two types of information in Web questionnaires raises the issue of how they interact to shape respondents' understanding of the questions.

One danger is that the visual information distracts respondents, leading to misinterpretation of the questions or other response problems. A second potential problem is that, when the visual information is taken as an essential part of the survey instrument (as opposed to an embellishment that can safely be ignored), it may affect the interpretation of the verbal information in the question text. Another potentially misleading feature of visual images is that they tend to be more concrete than words, which can be used to convey both concrete and abstract ideas (David, 1998). For this reason, visual images may not be good for illustrating general classes of items. But images can also be very complex, containing many different elements, potentially leading to many possible interpretations. The final potential problem involves the relation between the verbal and visual information; the two can be consistent or inconsistent with each other. When the two are consistent, *accentuation* effects can occur (e.g., Krueger and Rothbart, 1990), with the visual information reinforcing the verbal information and sometimes leading to more extreme interpretations. However, when the verbal and visual information are not congruent, *interference* effects can occur, with confused or slow responses the result.

Visual enhancements are widely used in Web surveys, both because they are easy to implement and because Web survey designers believe that aesthetically pleasing sites are needed to motivate respondents to complete the survey. But the impact of these design choices on measurement error is largely unexplored. Together with Roger Tourangeau, Fred Conrad and others, I am embarked on a program of research to explore these issues in detail.

In addition, while the task of designing or developing a Web survey rich in visual features may be relatively simple, the effort of ensuring that all respondents see the survey in the same way and are able to navigate, view and complete the survey in a consistent fashion should not be underestimated.

One of the basic tenets of survey measurement is the notion of standardization. Survey questionnaires or instruments are designed to present a standard stimulus to all respondents, and interviewers are trained, monitored and supervised in order to ensure standardized oral delivery. The Web offers much less control for the survey designer and, in fact, places such control in the hands of the respondent (explicitly or otherwise). Thus, variations in modem speed, computer operating system, Internet service provider (ISP), browser type and version, availability of plug-ins, and a host of hardware and software settings under the control of the individual user may all affect the way the survey instrument is received and viewed by the respondent. Three brief examples may serve to illustrate this point:

- Users have a variety of choices of security settings in browsers. One of these is to alert the user whenever data are transmitted to an insecure server. Many Web surveys do not make use of secure servers, and this may mean that respondents need to respond to several dialogue boxes in order to complete a single question.
- The size of the browser window (full screen versus reduced-size window) is under user control, as are font size settings (e.g., both small vs. large font selections in the Windows display settings, and font size adjustments within the browser). In addition, monitor resolution (e.g.,

800×600 or 1024×768) is not consistent for all respondents. Together these variations may results in some parts of the question not being visible to some respondents while fully visible to others, in uneven columns in tables, or in other variations in the presentation of questions across respondents.

- The user has control over the colours used for the browser (both background and foreground). If the designer is not careful, and does not explicitly override these settings, the survey may be unreadable or the colour enhancements may not have their intended effects.

There are many other user-defined options and variations in systems that may change the survey experience across respondents. Table 3 lists some of characteristics of browsers and operating systems, and shows the diversity of systems and settings that users bring to Web surveys. Several surveys are designed to require plug-ins (e.g., video or sound players, viewers for proprietary images, Java applets, etc.) that in my experience rarely work as intended. In addition, there are a host of emerging wireless or portable Internet access technologies that will only serve to make the design of Web surveys even more challenging.

In terms of colour, all browsers are *not* created equal. What may work on one browser may be unreadable on another. There are ways to avoid these problems (for example, using browser-safe colours), but again this requires careful attention to design details that are often overlooked in the frenetic world of Web surveys where speed and cost appear to be key driving factors.

Thus, while the Web offers great promise for creative and innovative survey design and measurement, great care must be taken to ensure that respondents receive, view, and interact with the instrument in a similar way.

Thus far I have been talking generically about measurement issues in Web surveys. However, in the same way that there are many varieties of Web surveys in terms of sampling and selection, so too are there different flavours in terms of instrument design. These can generally be grouped into two main approaches: static (or scrollable) Web surveys and dynamic (or interactive) Web surveys. As suggested earlier, Web surveys have developed out of two traditions—self-administered surveys (primarily mail) and computer-assisted interviewing (primarily CASI). These two pedigrees have led to very different views on the design of Web survey instruments.

The view of Web surveys as electronic versions of mail surveys is championed by Dillman (2000), among others. Web surveys are typically designed as a single HTML form, with the users scrolling through the instrument much as they would a paper questionnaire. Skips are not automated, and edit or range checks are not present. Until the respondent presses the "submit" button at the end of the survey, no data are transmitted to the server, and changes can be made to answers at will. Respondents can choose to leave items unanswered, and can scroll back to review (and change) previous answers.

Web surveys developed out of the computer-assisted interviewing tradition are more interactive in nature. One or more questions is presented on a screen, and responses are transmitted to the server after every question or set of questions. This permits automated skips, range and error checks, and all the other functionality of CATI or CAPI instruments. Respondents can be permitted to page back through earlier answers, or this can be prevented. Similarly, respondents can be permitted to skip an item or be forced to provide a response to every question. Feedback can be provided to respondents depending on the answers they provided. Finally, this approach permits randomisation, one of the key benefits of Web-based data collection.

I do not mean to imply that one approach is superior to the other for all applications. Both designs have their uses, depending on the length, complexity and content of the survey, and the target audience. However, the two approaches have different implications for data quality. Furthermore, there is increasing variation within these basic themes. For example, one can embed Java applications into an HTML form (or use dynamic HTML) to provide immediate feedback or use continuous measures like slider bars. Similarly, pop-up surveys often use JavaScript to present the user with a separate window containing the survey items. Each of these approaches has different design implications. So, even here, we are not talking of Web surveys as a single technique or approach, but the particular design choices one makes could affect both non-response error and measurement error.

6. Summary and Conclusions

For many market research applications, the sampling, coverage, non-response and measurement problems may be of little consequence. The researcher or client may not be interested in broad generalizations to a larger population, and may simply want to gain a sense of general reaction to a product or advertisement from as large and diverse a group as possible in a short time and with relatively little cost. This may well be sufficient to inform decisions about packaging, product placement, pricing, and so on. For such uses, Web surveys may well be an ideal method of data collection. One area where this may be especially true is in advertising research. The capacity of the Web to deliver full colour images, sound, video, and other rich audio-visual stimuli to large numbers of respondents in a controlled (i.e., randomised, timed, ordered, standardized) fashion is unparalleled. Similarly, the Web offers a great tool for experimentation where the focus is on randomisation rather than representation.

On the other hand, I'm sure most of us would be concerned if major public policy decisions were based on similarly low-rigor methods. Estimates of unemployment, welfare recipiency and program participation, crime victimization, consumer expenditures, and a host of other government survey programs would be ill served by data based on such surveys. For such policy research, most Web surveys cannot compete with traditional interviewer-administered approaches in terms of accuracy and the richness of the information that can be obtained. Yet, for such sectors of the industry, Web surveys may well serve a supplemental or supporting role, particularly in mixed-mode applications or surveys of specialized populations (e.g., business surveys).

As I stated at the outset, it is important to make the standards of comparison explicit. Several industry bodies, including ESOMAR, the Council of American Survey Research Organizations (CASRO) and the National Council on Public Polls (NCPP) have issued guidelines or standards for methodological disclosure of Web surveys. If adhered to, these may serve as a basis for evaluating the claims made by the purveyors of Web surveys.

It is the self-selected survey purporting to provide projectable, scientific data that is the most insidious form of Web survey and present the greatest potential threat to the survey industry. This is made more of a concern because of the relatively low statistical literacy on the part of the lay public and policy-makers alike. Few can distinguish good surveys from bad—indeed many have trouble distinguishing survey data from those gathered by other less-rigorous methods. Without such knowledge, all data are valued equally. Under such circumstances, size (i.e., the number of cases) becomes the only accessible measure of quality, and price becomes the only measure of value, a dangerous state of affairs indeed.

I want to make it clear that I am no enemy of Web surveys. Indeed, I have embraced the new method and am using Web surveys in many different applications. The key for me is to be cognizant of the limitations of the method and to work within such boundaries. Web surveys have a legitimate place in the world of survey research, and have a very bright future. What does concern me, however, are the hyperbolic claims for Web surveys purporting to offer what they presently can not—that is, probability samples of the general population. Our focus should be on learning what the limits of the new methods are and finding ways to capitalize on their strengths.

The broad rubric "Web survey" has already come to include a vast array of different designs and approaches, not only in terms of sampling and selection procedures, but also in terms of design and implementation. As with any other survey technique, it is not the tool *per se* that is flawed, but the way it is used that determines the likely quality of the outcome. Thus, to dismiss all Web surveys as unscientific is much too sweeping a derogation. In same measure, to claim that Web surveys will soon replace all other modes of survey data collection is equally overstated. As with all other methods, Web surveys represent the good, the bad, and the ugly of survey research. The hope of those who believe that survey research is a valuable scientific method is that in time the good will prevail.

Appendix

Table 1. Percent Online by Selected Demographic Characteristics, November/December 2000

Characteristic	Percent online
Age	
18-29	75%
30-49	65%
50-64	51%
65+	15%
Income	
Under $30,000	38%
$30,000-$50,000	64%
$50,000-$70,000	72%
$70,000 +	82%
Education	
High school or less	37%
Some college	71%
College degree or more	82%

Source: Pew Internet Project: Internet Tracking Report (February, 2001)

Table 2. Example Web Survey Response Rates

Source	Target Population, Topic and Design	Response Rate by Mode
Kwak and Radler (2000)	Univ. of Wisconsin students; 1999 survey on campus computing; n=1000 per mode	Web: 27.4% Mail: 41.9%
Radler (2000)	Univ. of Wisconsin students; 2000 survey on campus computing; n=1000 per mode	Web: 28.0% Mail: 52.6%
Guterbock et al. (2000)	University of Virginia students; survey on university computing; mode comparison	Web: 36.8% Mail: 47.6%
Bason (2000)	Univ. of Georgia students; survey on drug and alcohol use; mode comparison, n=200 per mode	Web: 15.5% Mail: 27.7% IVR: 17.4% Phone: 23.9%
Kennedy et al. (2000)	National Survey of Student Engagement pilot; students at multiple U.S. universities	Web only: 38.5% Mail with Web option: 42.7% (7.8% via Web, 34.9% via mail)
Jones and Pitt (1999)	Staff at 10 English universities; mode comparison	Web: 19% E-mail: 24% Mail: 72%
Weible and Wallace (1998)	MIS faculty; mode comparison, n=200 per group	Web: 26% E-mail: 24% Mail: 35% Fax: 25%
Elig, Quigley, and Hoover (2000)	Survey of U.S. military personnel; mode comparison	Web, with paper option: 7% Mail, with Web option: 14% Mail, no mention of Web: 12%
Quigley and Reimer (2000)	Survey of U.S. military personnel; mode comparison	Web with paper option: 37% (27% via Web, 10% via paper) Mail with Web option: 42% (10% via Web, 32% via mail) Mail only: 40%

Source	Target Population, Topic and Design	Response Rate by Mode
Ramirez, Sharp, and Foster (2000)	Survey of full-time employees of the U.S. General Accounting Office; e-mail invitation for Web survey, with option to do paper if requested	Web with mail option: 75.3% Web, 11.6% paper; 87% total
Comley (1998)	Sample of readers on an Internet magazine; all those with e-mail addresses assigned to Web options, others to mail	Mail with Web option: 17% (1% via Web, 16% via mail) E-mail 13.5% Mail only: 15.4%
Zhang (2000)	Web survey of researchers in Library and Information Science; e-mail invitation sent to 201 authors whose e-mail address was available	Total: 79.6% (62.2% via Web, 15.4% via mail or fax)
Crawford, Couper, and Lamias (2001)	University of Michigan students; Web survey on affirmative action	Total Web: 34.5% (24.4% completes, 10.1% partials)
Couper, Traugott and Lamias (2001)	University of Michigan students; Web survey on affirmative action	Total Web: 47.1% (41.5% completes, 5.6% partials)
Univ. of Colorado (2000)	Survey of all seniors; mailed questionnaire with Web option	Total: 60% (13% via Web, 48% via mail)
Hill (2001)	Nurses; sent letter with URL and PIN	Web: 6%
Frey (2000)	U.K. mailbase and list from two New Zealand universities; survey on moral intensity; e-mail invitation	Web: 9.8%
McCulloch et al. (2000)	Web survey of AAPOR members; sent letter with URL and password	Web: 34%

Table 3. Key Browser Statistics, January 2001

Characteristic	Percent of browsers
Screen resolution:	
800x600	54%
1024x768	30%
640x480	7%
Other, unknown	9%
Javascript and Java:	
Javascript enabled	81%
Java enabled	78%

Characteristic	**Percent of browsers**
Browsers:	
MSIE 5.x	72%
MSIE 4.x	12%
Netscape 4.x	10%
Other	6%
Operating system:	
Win 98	68.92%
Win 95	13.08%
Win NT	6.67%
Win 2000	4.54%
Unknown	3.55%
Mac	2.10%
WebTV, Linux, Unix	0.98%

Source: http://www.thecounter.com

References

Ahlhauser, B. (1999), "Introductory Notes on Web Interviewing." *Quirk's Marketing Research Review*, Article 0509, July. <www.quirks.com>

Bason, J.J. (2000), "Comparison of Telephone, Mail, Web, and IVR Surveys." Paper presented at the annual meeting of the American Association for Public Opinion Research, Portland, OR, May.

Black, G.S. (1998), "The Advent of Internet Research: A Replacement Technology." Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Louis, May.

Black, G.S. and Terhanian, G. (1998), "Using the Internet for Election Forecasting." *The Polling Report*, October 26th.

Cohen, M. (1999), Presentation to the Council of American Survey Research Organizations, November 3rd.

Comley, P. (1998), "The Use of the Internet for Opinion Polls." Virtual Surveys, Ltd., unpublished paper.

Comley, P. (2000), "Pop-Up Surveys. What Works, What Doesn't Work and What Will Work in the Future." Paper Presented at the ESOMAR Net Effects Internet Conference, Dublin, April.

Couper, M.P. (2001), "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly*, 64 (4), 464-494.

Couper, M.P. and Nicholls II, W.L. (1998), "The History and Development of Computer Assisted Survey Information Collection." In Couper, M.P., et al. (eds.) (1998), *Computer Assisted Survey Information Collection*. New York: Wiley.

Couper, M.P., Tourangeau, R., and Steiger, D.M. (2001), "Social Presence in Web Surveys." Paper accepted for presentation at CHI '01, Conference on Human Factors in Computing Systems, Seattle, April.

Couper, M.P., Traugott, M., and Lamias, M. (2001), "Web Survey Design and Administration." *Public Opinion Quarterly*, forthcoming.

Crawford, S., Couper, M.P., and Lamias, M. (2001), "Web Surveys: Perceptions of Burden." *Social Science Computer Review*, forthcoming.

David, P. (1998), "News Concreteness and Visual-Verbal Association; Do News Pictures Narrow the Gap Between Concrete and Abstract News?" *Human Communication Research*, 25 (2): 180-201.

Dillman, D.A. (2000), *Mail and Internet Surveys; The Tailored Design Method*. New York: Wiley.

Elig, T.W., Quigley, B., and Hoover, E.C. (2000), "Response Rate Effects of Making Web or Paper the Primary Mode." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Frey, B.F. (2000), "Investigating Moral Intensity with the World-Wide Web: A Look at Participant Reactions and a Comparison of Methods." *Behavior Research Methods, Instruments, and Computers*, 32 (3): 423-431.

Groves, R.M. (1989), *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. and Couper, M.P. (1998), *Non-response in Household Interview Surveys*. New York: Wiley.

Guterbock, T.M., Meekins, B.J., Weaver, A.C., and Fries, J.C. (2000), "Web Versus Paper: a Mode Experiment in a Survey of University Computing." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Hill (2001), AAPORNET submission, March 1st.

Jones, R. and Pitt, N. (1999), "Health Surveys in the Workplace: Comparison of Post, Email and World Wide Web Methods." *Occupational Medicine*, 49 (8):556-558.

Kehoe, C.M. and Pitkow, J. (1996), "Surveying the Territory: GVU's Five WWW User Surveys." *The World Wide Web Journal*, 1 (3): 77_84.

Kennedy, J.M., Kuh, G., Li, S., Hayek, J., Inghram, J., Bannister, N., and Segar, K. (2000), "Web and Mail Survey – Comparison on a Large-Scale Project." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Krueger, J. and Rothbart, M. (1999), "Contrast and Accentuation Effects on Category Learning." *Journal of Personality and Social Psychology*, 59 (4): 651-664.

Kwak, N. and Radler, B.T. (2000), "Using the Web for Public Opinion Research: A Comparative Analysis Between Data Collected via Mail and the Web." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

MacElroy, B. (2000), "Variables Influencing Dropout Rates in Web-Based Surveys." *Quirk's Marketing Research Review*, July/August (<www.modalis.com>)

McCulloch, C., Nieves, J., Nyiel, Z., and Tenore, V. (2000), "Attitudes Toward the State of the Survey Research Industry." Paper presented at the annual meeting of the American Association for Public Opinion Research, Portland, OR, May.

McLaughlin, T. (2000), "Customer Database Research: Guidelines for Complete and Ethical Data Collection." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Mitofsky, W.J. (1999), "Pollsters.com." *Public Perspective*, June/July: 24-26.

National Telecommunications and Information Administration (NTIA) (2000), *Falling Through the Net: Toward Digital Inclusion*. Washington, DC: U.S. Department of Commerce.

O'Muircheartaigh, C. (1997), "Measurement Errors in Surveys: A Historical Perspective." In Lyberg. L.E., Biemer, P.P., Collins, M., de Leeuw, E.D., Dippo, C., Schwarz, N., and Trewin, D. (eds.), *Survey Measurement and Process Quality*. New York: Wiley.

Paivio, A. (1991), "Dual Coding Theory–-Retrospect and Current Status." *Canadian Journal of Psychology*, 45: 255-287.

Quigley, B., Riemer, R.A., Cruzen, D.A., and Rosen, S. (2000), "Internet Versus Paper Survey Administration: Preliminary Findings on Response Rates." Paper presented at the 42nd annual conference of the International Military Testing Association, Edinburgh, Scotland, November.

Ramirez, C., Sharp, K., and Foster, L. (2000), "Mode Effects in an Internet/Paper Survey of Employees." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H.-J., and Berry, S. (2000), "Comparing Random Digit Dial with Web Surveys: The Case of Health Care Consumers in California." Santa Monica, CA: RAND, unpublished paper.

Schmidt, J. (2000), " 'Bill of Rights of the Digital Consumer': The Importance of Protecting the Consumer's Right to Online Privacy." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Terhanian, G. (1999), "Understanding Online Research; Lessons from the Harris Poll Online." Paper presented at the annual conference of the American Association for Public Opinion Research, St. Petersburg Beach, FL, May.

Terhanian, G. (2000), "How To Produce Credible, Trustworthy Information Through Internet-Based Survey Research." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Tuten, T.L., Bosnjak, M., and Bandilla, W. (2000), "Banner-Advertised Web Surveys." *Marketing Research*, 11 (4): 17-21.

University of Colorado (2000), online report of student survey.

Weible, R. and Wallace, J. (1998), "Cyber Research: The Impact of the Internet on Data Collection." *Marketing Research*, 10 (3): 19-24, 31.

Witte, J.C., Amoroso, L.M., and Howard, P.E.N. (2000), "Method and Representation in Internet-Based Survey Tools – Mobility, Community, and Cultural Identity in Survey2000." *Social Science Computer Review*, 18 (2): 179-195.

Yoshimura, O. and Ohsumi, N. (1999), "Some Experimental Surveys on the WWW Environments in Japan." Paper presented at the International Symposium on New Techniques of Statistical Data Acquisition.

Zhang, Y. (1999), "Using the Internet for Survey Research: a Case Study." *Journal of the American Society for Information Science*, 51 (1): 57-68.

About the Author

Mick P. Couper is a senior associate research scientist in the Institute for Social Research's Survey Research Center and adjunct associate professor in the Department of Sociology at the University of Michigan, and a research associate professor in the Joint Program in survey Methodology.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

The Internet as a Mode of Data Collection in Government Social Surveys: Issues and Investigations

John Flatley

Abstract

There is growing interest in the Internet as a tool for survey data collection. National statistical institutes are eager to examine how the Internet can be utilised to speed the process of collecting data; to reduce the costs associated with collection and processing, and to make data accessible more widely and more timely. The focus of this paper is web surveys in the context of government social surveys of the general population. The traditional concerns of survey researchers in government, such as reduction of bias through probability sampling, high response rates, and question reliability and validity raise fundamental questions for web surveys. This paper examines these issues and considers the implications raised.

Keywords

Internet access, government social surveys, web surveys, web CASI

1. Introduction

For those in the statistics business, web technologies offer the potential to speed the process of collecting data; to reduce the costs associated with collection and processing, and to make data accessible more widely and more timely. It is not surprising, therefore, that national statistical institutes (NSIs), such as the Office for National Statistics (ONS), are racing to embrace the Internet to enhance the service they deliver to government, other customers and to citizens. This is the business imperative. In addition, governments have added political impetus. For example, here in the United Kingdom (UK) central government has set a target for all public services to be available on-line by 2005. Similar, but more ambitious, targets have been set in Australia (2001) and Canada (2004). As the number with Internet access grows and people become accustomed to using it, there is likely to be increasing demand from the providers and users of official statistics to use the web for such purposes.

Social Survey Division (SSD) of the ONS is examining the feasibility of using the web for the surveys that it undertakes. Nationally representative surveys of the private household population carried out on behalf of ONS or other government departments, forms the core business of SSD. Such surveys employ random probability sample designs and are voluntary. Surveys undertaken by SSD tend to be undertaken with trained interviewers using computer assisted interviewing (CAI). Computer assisted personal interviewing (CAPI) is often the single mode of data collection. There are examples of surveys that combine CAPI with another mode. For example, the Labour Force Survey (LFS), which is a panel survey, uses CAPI for first interview and offers a computer assisted telephone interviewing (CATI) option for subsequent interviews (there are five interviews in total). On both the Family Expenditure Survey (FES) and National Travel Survey (NTS), following an initial CAPI interview, respondents are asked to self-complete a paper diary. SSD is also commissioned to sample other populations, such as school pupils, in which paper and pencil interviewing continue to have a role though interest in using the web for such surveys may increase.

The characteristics of government social surveys of the general population do not readily lend themselves to web data collection. Social surveys of the private household population tend to have the following requirements:

- random probability sampling
- long and complex questionnaires, e.g. interviews of 1 hour or more are typical
- high response rates, e.g. 80% plus are a common requirement

These features contrast with common practice on web surveys and raises questions about the potential use of the Internet for general population surveys. ONS, like other NSIs, has begun to consider the issues that are raised by the possible collection of data via the web on social surveys of the general population. This paper draws on initial work in which, following a literature review and discussions with others already working in the field[6], we have identified a number of key issues. We start by considering the issue of population coverage and sampling and present the latest government statistics on Internet penetration.

2. Population coverage and sampling

Use of Internet

The National Statistics Omnibus survey started to include questions on the Internet on a regular basis in July 2000 and have been asked once every three months. The Omnibus Survey has a random probability sample design such that it yields results that are nationally representative of all *adults* in Great Britain.

According to the January 2001 Omnibus Survey, an estimated 23 million adults in Britain have used the Internet at some time either at their own home or elsewhere, such as at work or another person's home (ONS, March 2001). This is equivalent to half (51%) of the adult population in Great Britain and compares with a figure of 45% in July 2000 (ONS, September 2000).

Eight in ten of adults who had ever used the Internet by the time of the latest survey (January 2001) had accessed it in the month prior to the interview. This represents some 40% of Britain's adult population. The Omnibus Survey shows that a higher proportion of men (57%) had used the Internet than women (45%). This gender gap is consistent by age.

[6] This paper draws on the work already undertaken by other NSIs, especially in the United States and Canada, and the author is grateful to those who have so willing shared their experience. In particular, I am grateful to Howard Kanarek and Barabara Sedivi Gaul of the U.S. Bureau of the Census and other participants of the Web Surveys session at the recent FEDCASIC workshop in Washington DC for their helpful insights.

As expected, the proportion using the Internet declines with age. For example, among those aged 65-74 years 15% had used the Internet and just 6% among those aged 75 years and over. Overall two-thirds of all adults who had ever accessed the Internet were aged below 45 years old. Between July 2000 and January 2001 the largest increase in Internet usage was recorded for the youngest age group (up from 69% to 85%). This contrasted most strongly with those aged 75 years and over, among whom there was no growth recorded (6% at both times).

Home Internet access

Another source, the FES, can provide nationally representative estimates of the number of *households* with home Internet access. Figures for the FES refer to the whole of the UK. A question on home Internet access has been included on the FES since April 1998. Results on Internet access are produced on a quarterly basis and the latest figures relate to the final quarter of the year 2000 (October-December). Up to March 2000 this question was asked only for households who had already said that they had a home computer. As new methods of accessing the Internet have begun to be taken up the levels of Internet access in the first quarter of 2000 may be understated. From April 2000 all households have been asked about home access to the Internet and if they have it, the means of access.

In the first quarter in which such questions were included (April-June 1998), 9% of UK households were estimated to have home Internet access. This is equivalent to some 2.2 million households. Since the first quarter of 1999, the survey has recorded a steady increase in the proportion of households with access to the Internet at home. The latest estimates, for October-December 2000, are that 8.6 million households now have home access. This is equivalent to one in three UK households (35%) and is a four-fold increase on the comparable figure in 1998.

Analysis of the latest quarter shows that the vast majority of those with home Internet access (98%) use a home computer to access the Internet. Take-up of new technologies, such as WAP phones and digital televisions has yet to make a large impact. Over time the growth in the use of these new technologies may be the driving force behind increases in home Internet access. With the planned withdrawal of analogue transmitters by the end of this decade it is possible that nearly all UK households will have access to digital televisions, and possibly as a by-product home Internet access. This may have implications for the web survey designer. Currently such forms of web access do not have the same functionality as computers and few of the existing web surveys would run on such a platform. Another implication relates to possible differences in the profile of users. As yet there is little research on these issues.

The digital divide: unequal access to the Internet

The FES demonstrates that the proportion of households with home access to the Internet varies by a range of social and demographic factors. For example, FES data show that there is a strong relationship between household income and home Internet access. Over the financial year 1999-2000, the percentage of households with access to the net at home was lowest amongst households in the four lowest deciles (ranging between 3% and 6%). In both the fifth and sixth decile groups 15% of households had home Internet access. Amongst the remaining deciles access increased steadily with income, for example rising from 22% among those in the seventh decile to 48% for those in the highest decile (ONS, July 2000).

On the basis of two years data it is difficult to comment on the extent to which there is evidence of the narrowing of the gap between lower and higher income groups. In the last two years there was growth in home Internet access across all income decile groups and the pattern of unequal access was broadly similar.

Analysis of the January 2001 Omnibus Survey has shown a clear relationship between usage and social class. Here social class is based on the occupation, or previous occupation, of the household reference person (the person in the respondent's household whose name the accommodation is owned or rented, or if there two or more such people, the one with the highest personal income). For example, the proportion of adults who had used the Internet was higher than average (51%) among those living in households headed by a person in Social Class I, professionals and managers and Social Class II, Intermediate non-manual (78% and 65% respectively). This contrasted with those whose household head were in one of the manual social classes, where the proportions were as follows: 37% for those in Social Class III manual, 33% for those in Social Class IV partly skilled manual and 27% for those in Social Class V unskilled manual (ONS, March 2001).

Analysis of the 1999-2000 FES has shown households with the highest levels of access in 1999-2000 were those comprising couples with children (31% of couples with one child and 35% of couples with two or more children). This compared with an average figure of 19% for the whole of the UK at that time. Lower proportion of lone adult households had Internet access (7% for those with one child and 11% for those with two or more children). It is interesting to note the general pattern that, irrespective of the number of adults, rates of access were higher for households with two or more children compared with those containing one child. As expected, the lowest rates of access were found for households comprising people who were retired and living on their own or as couples (1% and 5% respectively) (ONS, July 2000).

While access to the Internet remains at a relatively low level, there is no prospect of the web replacing other modes of interviewing on high quality general population surveys that require nationally representative surveys. Even if access was to reach a near universal level, such as is the case with telephones (according to the General Household Survey, 96% of households have one), there is another challenge: the issue of sampling.

Sampling issues

Random probability sampling requires that each member of the population of interest has a known (non-zero) chance of being selected for the survey. In the UK, there is not a central register of the population, as is the case in some Scandinavian countries, to act as a sampling frame. For the population of adults living in private households, however, there are proxies such as the Postcode Address File (PAF) and Electoral Register (ER). In SSD, most of our general population survey samples are drawn from the PAF. Most of our surveys are carried out by our field force of trained interviewers using computer assisted interviewing (CAI) with laptops. Selected addresses are visited by interviewers to establish whether or not the address listed contains a household that is eligible for the survey and, if so, to attempt to persuade potential respondents to participate in the survey. Surveys that use face-to-face interviewing tend to find that around 11-12% of addresses listed in the PAF are found to be ineligible[7]. Some surveys over or under sample particular groups but because the chance of selection is known weighting strategies can be devised to adjust for known survey bias that would otherwise result.

Currently, there does not exist a single listing of people with access to the Internet. Email lists may exist for membership associations, such as the Association for Survey Computing (ASC) or the Social Research Association (SRA), and for employees of businesses and other organisations. If these are reasonably comprehensive and up-to-date, such lists could be used to invite all, or a sample of, members to respond to a web survey. No such list of the general population exists here in the UK; nor is one likely in the foreseeable future.

Even if a list were to become available for sampling purposes, we would have difficulty in ensuring that each sampled individual had a known chance of being selected because individuals may possess more than one email address at any one time. Some people with web access may not use email, even if they were required to set up an email account by an Internet Service Provider (ISP). Further, as people move from one ISP to another there is potential for email addresses once used to become dormant. We know little about the general public's use of email addresses and we don't have good answers to questions such as:

- What is the average number of email addresses that a household or individual maintains?
- What is the average life of an email address?
- What is the rate at which households or individuals acquire new email addresses?

In addition because of the great variety in the nomenclature of email addresses, there is little prospect of us being able to randomly generate email addresses in a manner that is possible, though not without problems, with random digit dialling.

3. Design and implementation issues

Screens and scrolling

There are two main approaches to questionnaire construction that have conventionally been used on web surveys. With a screen-based approach usually one question is displayed on each screen and each time a respondent answers a particular question the screen refreshes to display the next question. Scrolling refers to the situation when the whole of the questionnaire is available at one time by allowing the respondent to scroll up and down, using a scroll bar to the right of the screen. The choice of which method to use is often related to the way in which respondents are invited to complete the survey. Screen by screen presentations are often used with on-line surveys where there is the need for continuous contact between the respondent's computer and a web server. Off-line surveys often use a plain (Hypertext Mark-Up Language) HTML form which the respondent is able to download from a website or receive as an email attachment. The pros and cons of on-line and off-line approaches are discussed later. Here we're concerned simply with design issues.

[7] Addresses that are ineligible include those that are empty when an interviewer calls, that have been demolished or are derelict, are a business premise or used by an institution or are a second residence or holiday accommodation.

It has been argued that scrolling is more akin to the way in which people use computers and surf the web and for that reason should be preferred (Dillman, 2000). It is true that experienced web users are likely to find a screen-by-screen presentation more cumbersome. They may also be frustrated if they feel their freedom to navigate is constrained. This maybe compounded if the time delay between individual screens refreshing is perceived by respondents to be excessive.

Another advantage of scrolling is that respondents can more easily form a perception of the content and the length of the questionnaire before they begin their task. This makes scrolling more like conventional paper self-completion questionnaires where the respondent can easily flick through the document before answering and can form a judgement about how long it will take them to complete.

For these reasons, Dillman argues that scrolling should be preferred whenever web surveys are being used to replace paper self-completion methods. However, if web surveys are being used in a mixed mode environment, alongside CAPI or CATI, then the case for using a screen-by-screen construction is stronger. One could argue that it more akin to the way in which questions are 'served' and answered on these modes.

Off-the-shelf software is increasing the choices available to questionnaire developers and greater flexibility in questionnaire construction is becoming possible.

On-line versus off-line interviewing

An on-line survey is one in which the respondent accesses the questionnaire instrument on a website and completes it, via their web browser, while connected to the Internet. Respondents can be directed to the URL either through paper instructions, for example sent by regular mail or left by an interviewer, or in an email, possibly with a hyperlink.

On-line surveys need constant contact between the respondent's computer and the web server upon which the instrument is located. Thus one needs to consider whether or not this requirement is likely to be acceptable to respondents. For Intranet surveys, for example of employees of a business or other organisation, this is likely to be perfectly acceptable. One proviso is that the connection speed between respondent's computer and web server must be sufficient to ensure that there are not unacceptable delays in the processing of individual questions.

However, on-line methods raise a number of issues for general population surveys. The first of these concerns the length of time that a respondent is required to be on-line to complete the survey. This is likely to be of particular concern to those who pay by the minute for their time on-line. One could offer to compensate respondents for the cost of their time on-line but we need more research to find out whether or not this will be perceived as an adequate incentive for all respondents.

As outlined above, on-line surveys can be problematic if the time it takes for the response to one question being transmitted to the server and for the server to respond is deemed to be excessive. For general population surveys, where few households currently have high speed Internet access, this is a key concern. This suggests that on-line surveys are only feasible for general population surveys where there are few questions being asked and the duration of the interview is short.

In an off-line survey, respondents may initially go on-line to download a file, containing the survey instrument, to their local computer, completing the survey off-line before going back on-line to transmit the completed questionnaire back to the host. An alternative is for respondents to be sent a file, either by regular mail on CD or floppy disk or as an attachment to an email. Respondents need to install the file on their computer before answering the questionnaire. Once the questionnaire is completed, respondents can go on-line to return the completed questionnaire. There are other concerns, such as security or computer functionality, that may be relevant to a decision about on-line or off-line surveys. Respondents' competence in the use of the Internet may need to be higher to complete all the required tasks than is the case of on-line use.

Off-line surveys are preferred when the survey is longer and complex. However, off-line surveys can introduce other concerns. For example, the time that it takes for the questionnaire instrument to download to the respondent's computer may be perceived to be excessive. In addition, respondents may be unwilling to download files to their own hard drive worried that they may contain viruses or otherwise damage their computer. A possible alternative to downloading is to send respondents a file via regular mail, for example on CD or floppy disk, or as an email attachment. However, as with the download method this still requires some degree of trust in the integrity of the files being sent. It adds to the burden on respondents and for some maybe enough to dissuade them to participate. Off-line surveys are likely to work best when the files that need to be installed on a respondent's computer are relatively small. This means that there maybe less opportunity to build complex instruments, for example with extensive interactive edits and multi-media, if as a result respondents are required to download large computer files.

Plain HTML versus advanced programming

The simplest approach to designing web questionnaires is to provide a plain HTML form that respondents can access via their normal browser. However, plain HTML is limited. It is necessary to supplement HTML with the use of advanced programming languages, such as Java, to allow the use of interactive edits and routing and the addition of features such as pop-up help and progress bars.

One could argue that the possibility of including such features represents a radical step forward compared with what is possible with conventional paper self-completion forms. However, the use of advanced programming languages can be problematic since different browsers may be inconsistent in the way in which they interpret Java and JavaScript. Further, the addition of such features necessarily results in the file size of the instrument growing. This impacts on download times and increases the risk of non-response.

It is worth noting that usability tests performed by the US Census Bureau suggested that respondents rarely made use of on-line help. This was true irrespective of how it was presented, for example via buttons, hyperlinks or icons (Kanarek and Sedivi, 1999).

Security

Security is a major issue for government surveys. Respondents to official surveys need to be assured that the data that they give us will be kept secure. At the same time, we need to devise systems that protect the integrity of our own data from hackers (Ramos, Sedivi and Sweet, 1998 and Clayton and Werking, 1998).

The US Census Bureau has developed a three level security system including the use of encryption, authentication and a firewall. Encryption provides strong security of data on the Internet while it passes between respondent and the Bureau. Their encryption solution requires respondents to have version 4.0 or higher of either Netscape Communicator or Microsoft's Internet Explorer and enables strong 128-bit encryption. Once received at Census bureau, the data are protected by a firewall that guarantees that others from outside cannot have access to the data (Kanarek and Sedivi, 1999).

There is a tension between the strength of security and survey response. The experience of the US Census Bureau has been that the more stringent the security the lower is the response. In part, this appears to be related to technical issues, such as respondents not having the required browser version or encryption level. It is also possible that requiring the use of usernames and passwords to access a web survey may be problematic for some respondents (Sedivi, Nichols, Kanarek, 2000)

Data quality

One of the attractions of web based data collection, compared with paper self-completions, is the improvement in data quality that is offered with the use of CAI methods. If the technical issues noted above can be overcome, WEB CASI opens up the possibility for more complex questionnaires to be self-administered than is possible with paper based approaches. Complex routing can be incorporated in a way that is not feasible in a paper document. Computations and edits can be carried out at the time of the interview and inconsistent or improbable responses checked with the respondent. This is analogous of the move from pencil and paper methods to CAPI and CATI.

An important aspect of data quality is the reduction of measurement error. This arises when inaccurate answers are recorded for respondents. There are different sources of measurement error, such as poor question wording and differences arising from the mode in which the interview is conducted. It is possible that there will be mode effects arising from the switch from other modes to web CASI. For example, compared with CAPI or CATI, there is a difference between the way in which information is presented to respondents – a move from aural to visual presentation. The speed with which respondents read and respond to questions can't easily be controlled. Clearly more research is needed in this area to examine such questions.

Survey response

Currently, web surveys have tended to achieve lower response than mail surveys. The US Census Bureau has found no evidence from its trials, largely among businesses, that offering a web option in addition to other modes leads to an overall increase in response rate (Sedivi, Nichols and Kanarek, 2000). Part of this is likely to be accounted for technical problems, discussed earlier. Another possible explanation offered relates to respondents' concerns about privacy and confidentiality (Couper, 2000).

Response rates to self-completion surveys tend to be lower than to interviewer-administered surveys and there is every reason to expect that this will be the case for web surveys too. CAPI and CATI achieve higher response than mail surveys because of interviewer involvement in the initial recruitment of respondents and in sustaining respondent involvement. Interviewers can be good at persuading respondents to participate by explaining the value of the survey and the respondents' participation. During the interview itself, interviewers can develop a rapport with respondents that encourages them to continue. For conventional surveys, the proportion of respondents who refuse to complete an interview once started is low. This contrasts with web surveys. Opinion on the optimal length of web surveys is divided but anecdotal evidence suggests that 15-20 minutes is as much as one can expect. This compares with CAPI/CATI surveys that can range from 30-40 minutes up to one hour or more.

One interesting feature about the analysis of the profile of the on-line population, presented earlier, is the over-representation of younger adults. This is one sub-group that tends to be under-represented in random probability sample surveys of the general population. WEB CASI might be a way to increase response in this sub-group. However, there is evidence to suggest that it is not a good idea, in mixed mode surveys, to offer respondents a choice of response mode. For example, the US Census Bureau tested whether or not response was boosted by offering both mail and telephone options. Overall 5% of the sample responded by phone but the overall level of response (mail and phone combined) was the same as for when only a mail option was offered (Dillman, Clark and West, 1995).

Questionnaire designers need to be aware that the use of more advanced techniques may impact on respondents' ability to respond. The more complex the questionnaire instrument is, and the more that elaborate programming is embedded within it, the greater will be the size of the resulting file that respondents will need to download to run the questionnaire. This will impact on the time it takes to download the questionnaire and is very likely to impact on response.

Design principles

It has been argued that the same visual design principles that apply to the design of paper questionnaires apply for web surveys (Dillman, 2000). Respondents have the same requirement for information that is presented clearly and efficiently. Layout and design should aim for respondents to read every word of each question and in a prescribed order.

Dillman had developed a set of 14 design principles for web surveys. These are:

1. Introduce the Web questionnaire with a welcome screen that is motivational, emphasises the ease of responding, and instructs respondents about how to proceed to the next page
2. Provide a PIN number for limiting access only to people in the sample
3. Choose for the first question an item that is likely to be interesting to most respondents, easily answered, and fully visible on the welcome screen of the questionnaire
4. Present each question in a conventional format similar to that normally used on paper self-administered questionnaires
5. Restrain the use of colour so that figure/ground consistency and readability are maintained, navigational flow is unimpeded, and measurement properties of questions are maintained
6. Avoid differences in the visual appearance of questions that result from different screen configurations, operating systems, browsers, partial screen displays, and wrap-around text
7. Provide specific instructions on how to take each necessary computer action for responding to the questionnaire, and give other necessary instructions at the point where they are needed
8. Use drop-down boxes sparingly, consider the mode implications, and identify each with a "click-here" instruction
9. Do not require respondents to provide an answer to each question before being allowed to answer any subsequent ones
10. Provide skip directions in a way that encourages marking of answers and being able to click to the next applicable question
11. Construct web questionnaires so they scroll from question to question unless order effects are a major concern, or when telephone and Web survey results are being combined

12. When the number of answer choices exceeds the number that can be displayed in a single column on one screen, consider double-banking with an appropriate grouping device to link them together

13. Use graphical symbols or words that convey a sense of where the respondent is in the completion process, but avoid those that require significant increase in computer resources

14. Exercise restraint in the use of question structures that have known measurement problems on paper questionnaires, such as check-all-that apply and open-ended questions

Another view is that the web is a fundamentally different medium from paper and that much more research is required before we can determine optimal designs (Couper, 2000). Couper suggests that design issues interact with the type of web survey being conducted and the population which is being targeted. For example, an optimal design for a survey of teenagers may be quite different from that which is required for a survey of elderly people.

Dealing with the variety of computer hardware and software that may be used by respondents to access the survey also needs to be kept in mind. A range of factors, such as operating system, browser type and version and Internet connection speed may affect the functionality of the survey instrument. Most of these are outside our control. The result maybe that some respondents may not be able to access the questionnaire at all, others may find it slow to respond and/or the questionnaire does not display as intended.

4. Conclusion

This paper has been concerned with the use of the web for government *social surveys* of the general population. Such surveys are required to yield nationally representative estimates. The fact that the proportion of households with Internet access remains low and that the on-line population differs in important respects from the general population presents a major challenge. Issues of population coverage and sampling mean that it is not possible to achieve good population coverage with stand-alone web-surveys. This suggests that in the immediate future the most likely application of the web in official social surveys is as part of a mixed mode data collection strategy alongside others, such as mail, CAPI or CATI.

It is worth noting that, to date, it is with surveys of businesses that NSIs have made most progress with using the Internet as a mode of data collection. Rates of Internet access are relatively high (compared with households) and an increasing number are using the net to conduct their business. Most of the surveys that have offered a web option are relatively short (compared with social surveys) and self-completed. Here the web offers the potential to reduce costs[8], speed processing and improve data quality (e.g. with use of interactive edits).

A number of the issues raised in this paper are common to both surveys of businesses and social surveys of the general population. However, the challenges are greater for social surveys. For general population surveys that have self-completion elements there is growing interest in reporting via the web. Offering a web option for parts of the survey that are interviewer-administered raises even more

[8] Adding a web option alongside other modes is likely to increase survey costs in the short-term until the point is reached where the number of respondents using the web results in sufficient cost savings in data processing to cover the additional costs.

challenging questions, such as the effect on overall response rates and comparability with data collected by other modes.

NSIs are starting to grapple with the issues raised in this paper and the ONS has established a project (the *WEB CASI* project) within SSD to investigate these issues and take this work forward. Some of the points raised in this paper will be explored in a trial study during which former respondents to one of ONS' surveys, who reported having home Internet access at the time at which they were interviewed, will be approached to participate in a trial study. The results of our work will be shared within the social survey research community to contribute to the development of our knowledge in this area.

References

Couper MP (2000) Web Surveys: A Review of Issues and Approaches *Public Opinion Quarterly Volume 64*, American Association for Public Opinion Research.

Clayton RL and Werking GS (1998) Business Surveys of the future: the world wide web as a data collection methodology in Couper MP, Baker RP, Bethlehem J, Clark CZF, Martin J, Nicholls II W L and O'Reilly JM (eds) Computer Assisted Survey Information Collection, John Wiley and Sons.

Dillman D, Clark JR and West KK (1995) Influence of an invitation to answer by telephone on response rates to census questionnaires *Public Option Quarterly, Volume 58*, American Association for Public Opinion Research.

Dillman D (2000) Mail and Internet Surveys, John Wiley.

Kanarek H and Sedivi B (1999) Internet Data Collection at the U.S. Census Bureau, Paper to Federal Committee on Statistical Methodology conference.

Office for National Statistics (July 2000) First Release: Internet Access, 1st Quarter 2000.

Office for National Statistics (September 2000) First Release: Internet Access.

Office for National Statistics (March 2001) First Release: Internet Access.

Ramos M, Sedivi B and Sweet E M (1998) Computerised Self-Administered Questionnaires in Couper MP, Baker RP, Bethlehem J, Clark CZF, Martin J, Nicholls II W L and O'Reilly JM (eds) Computer Assisted Survey Information Collection, John Wiley and Sons.

Sedivi B, Nichols E and Kanarek H (2000) Web-based collection of economic data at the U.S. Census Bureau Paper presented to ICES II Conference.

About the Author

John Flatley is a Senior Social Survey Officer in the Social Survey Division of the Office for National Statistics. He is Manager of SSD's WEB CASI project that is examining the feasibility of utilising the Internet as a mode of data collection in government social surveys. For more information contact john.flatley@ons.gov.uk, telephone: from UK: 020 7533 5530; international 44 20 7533 5530.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

The Recruitment of Online Samples by CATI-Screening: Problems of Non-response

Marc Deutschmann & Frank Faulbaum

Abstract

It is well known that self-selection may negatively influence the quality of data collection in the World Wide Web. Survey results that are based on self-recruitment by respondents are generally biased.

A commonly applied strategy to circumvent or at least to reduce the effects of self-selection is to recruit random samples of Internet-users by CATI-screening. Within the screening respondents are asked whether they use the Internet and whether they are willing to give their E-Mail addresses. This procedure guaranties known selection probabilities for the respondents and the selection process can be better controlled.

Up to now, we know rather little about the effects of the usual CATI problems concerning non-response caused by refusals and non-availability of respondents on the quality of the online sample. In addition, the change of the collection method may lead to new forms of non-response like e.g. the readiness for a telephone interview but not for an online interview.

A study of the University of Duisburg investigates the conditions for participating in online surveys and analyses attributes of refusers on the basis of data collected by CATI. The paper discusses the advantages and disadvantages of CATI-screening as a method of recruitment of online samples.

Keywords

Non-response, Web survey, Internet use, CATI-Sreening

1. Introductory Remarks

Despite the many advantages associated with the introduction of new technologies into data collection (see e.g. Nicholls, Baker & Martin 1997), some of the “traditional” challenging problems jeopardising the quality of survey results remain. Most of them refer to one or the other of the different types of errors possibly occurring while conducting a survey (see Biemer & Trewin 1997; Groves 1989; Groves & Couper 1998). Recently, Couper discussed the different types of errors with respect to Web surveys (see Couper 2000). In his article, he furthermore proposed a classification of Web surveys with respect to whether they use probability-based methods or non-probability methods of sampling. Within the latter he distinguished three different types of Web surveys: Polls as entertainment, unrestricted self-selected surveys and volunteer opt-in panels. Within the former he distinguished intercept surveys, list-based samples, Web option in mixed-mode surveys, pre-recruited panels of Internet users and pre-recruited panels of full population (see Couper 2000, p. 477).

The type of survey we refer to in this paper shares most of the features of a pre-recruited panel of Internet users. The only feature that it doesn't share is the panel property; i.e. the respondents were not asked to participate repeatedly in a panel. But the basic form of recruitment is the same. In fact, we consider the form of recruitment of a sample of potential participants in a Web survey as used in this kind of survey to be the only acceptable alternative to the costly (but perhaps optimal) method of pre-recruited panels of full population, at least as far as sampling is concerned.

Pre-recruited panels of Internet users require the following stages:

- In order to reduce the effect of self-selection, probability sampling from a precisely defined population, which in turn requires the specification of a proper sampling frame (see Särndal, Swensson & Wretman 1992, p. 9) in order to avoid under-coverage.
- Collection of background information and the identification of those people having Internet access.
- Collection of e-mail addresses of those having Internet access.
- Sending of an e-mail request to participate in an Internet survey

At each stage specific types of non-response may occur. What the present paper is dealing with is

- the item non-response associated with questions concerning the e-mail address during an e-mail screening of potential participants in a later Web survey which in fact means a unit non-response for this later survey;
- the unit non-response associated with the later Web survey of those who gave their e-mail addresses during the screening..

More specifically, the principal aim of the study the results of which are presented below was to explore the differences between the group of Internet users with access to e-mail addresses and those actually participating in a subsequent Web survey after explicit invitation to participate. The mass of analyses of Internet use are concerned with the differences between Internet users and the rest of the population. But up to now there exists little information about differences between those who supply their e-mail addresses and participate in the later survey and those who supply their e-mail addresses and do not participate.

The screening of e-mail addresses was done by CATI. Consequently the difference between participants and non-participants is intimately related to the more general question, whether there are specific characteristics of non-response associated with the transition from the Telephone mode to the Internet mode which, of course, at the same time means a transition from interviewer-administration to self-administration. It is well known that motivation of respondents is especially important in self-administered surveys (see e.g. Dillman 1978, 2000). The Web survey following the screening of e-mail addresses dealt with the evaluation of the CATI interview; i.e. the participants had to fill out a Web questionnaire consisting of twenty questions about possible weaknesses of the CATI-questionnaire, the response categories, interviewer characteristics, the perceived time of the interview, etc. The design of the Web questionnaire attempted to respect the recommendations of Dillman (2000).

2. Study Design

The small non-response study referred to in this paper was designed in connection with a nation-wide CATI survey on media use and health caring information behavior. The total survey sample had a sample size of 2.026 persons of age 16 and above. Within this survey the additional opportunity was offered to recruit Internet users as participants in a subsequent Web survey.

As mentioned above, the collection method for the screening of Internet use and e-mail addresses was CATI. Thus, the first stage of the recruitment process had to consist in probability sampling of telephone numbers of private households which in turn required a suitable sampling frame. In Germany, RDD telephone surveys are not useful due to the non-uniform structure of telephone numbers. The latter have short area codes in bigger cities and longer codes in rural areas. The total number length varies between eight and eleven digits. The problem of creating a sampling frame for probability sampling with known inclusion probabilities for private households with telephone is not easy to solve, since it can be shown, that often used compromises between list-based procedures and random generation procedures like Randomized-Last-Digits (RLD) or Randomized-Plus-Digit also share the property of unknown inclusion probabilities (see Gabler & Haeder 1997).

Gabler & Haeder (1997; 1998) developed a practical sampling procedure for the sampling of telephone numbers in Germany, which they proved mathematically to guarantee random sampling of private households with telephone numbers, including those not mentioned in the official telephone directory. In the present study precisely this procedure was applied. The main feature also guaranteeing the practical applicability is that it generates random numbers only for those number blocks in which telephone numbers can be found. Given a selected household, a person of the population of interest in the household was in our case selected by the Last-birthday procedure.

About two weeks after the CATI interview, a request together with a link to the Web site with the Online questionnaire was sent to these respondents who supplied their e-mail address. The verification of the identity of the respondent to the Web questionnaire with those in the CATI interview was assured by the type of questionnaire. Since the Web questionnaire referred to the former CATI interview only those persons were able to fill it out who actually participated in the main CATI survey.

In order to get some information about the reasons of not supplying the e-mail address, the reasons were registered by the interviewers just after the CATI interview.

3. Results

As Table 1 shows, in the subsample of 2.026 respondents 1.045 (51.6%) reported that they use the Internet. Among these, 865 persons (42.7%) used both Internet and e-mail. The members of this group will subsequently called Internet users. 48.4% did not use the Internet. The use of e-mail addresses does, of course, not necessarily mean possession of an individual Internet address. E.g. a woman may use the e-mail of her husband or somebody may use the e-mail address of a colleague.

Table 1: Internet and e-mail use

		male	female	Total	
		%	%	Number	%
	Use Internet + e-mail	54.5	33.4	865	42.7
	Use Internet, but no e-mail	6.8	10.5	865	8.9
	Do not use Internet	38.6	56.1	981	48.4
Total		100.0	100.0	2,026	100.0

During the CATI interview the respondents who reported that they were Internet users in the above narrow sense of using Internet and using an e-mail address at the same time, were asked whether they were willing to participate in a later Web survey. 578 respondents (66.9% of the 865) agreed to participate. 497 respondents (57.5%) also agreed and were in the position to give their e-mail address (see Table 2).

Table 2: Ready to participate in the Web survey?

		male	female	Total	
		%	%	Number	%
	Ready and provided e-mail address	65.4	47.2	497	57.5
	Ready, but did not provide e-mail address	7.2	12.1	81	9.4
	Not ready	27.4	40.6	287	33.2
Total		100.0	100.0	865	100.0

The reasons for not supplying an e-mail address despite personal use are listed in Table 3. The main reason were the lack of time, lack of confidence in anonymity and data protection, lack of an individual e-mail address and rarity of Internet use.

Table 3: Reasons for not providing e-mail address

```
_____________________________________________________________________
                                                            Pct of
                                                         Responses
lack of time                                                  19.1
lack of confidence in anonymity and
data protection                                               13.3
lack of own e-mail address                                    11.7
don't know e-mail address                                     10.1
rarity of Internet use                                         8.8
already given enough information in CATI-interview             8.8
fear of getting included in advertising
campaigns                                                      4.2
lack of experience with the Internet                           3.4
costs of being online                                          1.1
other reasons or no reasons given                             19.6
_____________________________________________________________________
                                          377 responses      100.0
```

Table 4 shows that more than half of those who supplied their e-mail address actually participated in the Web survey.

Table 4: Participation in Web survey

		Number	%
	Yes	262	30.3
	No, but provided e-mail address	235	27.2
	No	368	42.5
Total		865	100.0

Significant differences were found between participation rates of males and females. As table 5 shows, the overwhelming proportion of actual participants in the Web survey were males. This proportion was much higher than in the group of Internet users in general. The third column contains all respondents in the CATI screening who reported to use the Internet, but didn't give their e-mail address.

Table 5: Participation in Web survey by Gender

		Participation in Web survey			Total	
		Yes	No, but provided e-mail address	No		
		%	%	%	Number	%
Gender	male	67.9	59.6	45.7	486	56.2
	female	32.1	40.4	54.3	379	43.8
Total		100.0	100.0	100.0	865	100.0

No differences in participation rates could be found with respect to age and education. The corresponding results are not shown here. An important, but not very surprising result is that an effect of the frequency of Internet use could be found. Table 6 shows that more than half of the participants in the Web survey (57.6%) reported a daily use of the Internet, while only slightly over 36% of those not participating reported this kind of behaviour.

Table 6: Frequency of Internet use

		Participation in Web survey			Total	
		Yes	No, but provided e-mail address	No		
		%	%	%	Number	%
Frequency of Internet use	daily	57.6	47.2	36.4	396	45.8
	more than once per week	35.9	43.0	41.0	346	40.0
	once per week or less	6.5	9.8	22.6	123	14.2
Total		100.0	100.0	100.0	865	100.0

Another important variable influencing the participation behaviour was the amount of past experience with Web surveys. Within the group of participants about 46% had already participated in an Internet survey, compared with only 14% in the group of refusers (see Table 7).

Table 7: Participation in Web survey before (Queston in CATI interview)

		Participation in our Web survey			Total	
		Yes	No, but provided e-mail address	No		
		%	%	%	Number	%
Did you ever participate in a Web survey before?	Yes	45.8	25.1	13.9	230	26.6
	No	54.2	74.9	86.1	635	73.4
Total		100.0	100.0	100.0	865	100.0

Finally, some other Online activities were found which differentiate between the Web survey participants and the non-participants. These activities are listed in Table 8.

Table 8: Other Online activities

	Participation in Web survey			
Have you ever done…? ("Yes" in Percent)	Yes	No, but provided e-mail address	No	Total (Internet users)
Online shopping	62.0	56.5	37.5	50.6
Online trading	20.4	16.1	12.3	15.9
Home banking	54.1	43.9	36.9	44.3
Reading magazines online	62.7	53.8	46.5	53.7
Downloading program or music files	72.2	71.3	54.2	64.6
Participating in mailing lists	27.1	25.1	12.9	20.8
Participating in chat rooms	55.3	59.6	46.5	52.9

803 valid cases

4. Conclusions

The central topic of this paper is the differences between CATI-screened participants in a German Web survey and CATI-screened non-participants with respect to a set of selected variables. As one of the key problems associated with this kind of procedure, Couper (2000, p. 489) mentioned the low initial response rate to the recruitment interview and the low number of those who subsequently agree to participate. The present results show that the rate of those who supplied their e-mail-addresses was quite high (more than 50%). Other recent German studies the results of which have not yet been published show similar results. It is still open whether growing experiences with the Web survey, will lead to an increase or a decrease of this rate in the future. Though the results of this study clearly show that familiarity with the net, especially with past Web surveys, increases or decreases of the readiness to participate in the survey, future developments are not easy to predict. These may depend upon whether people can be convinced that data protection can be guaranteed. The data of this study show reveal a certain tendency not to supply the e-mail address because of doubts regarding anonymity.

Nevertheless, the results indicate at least for Germany that if in the future more and more people not only use the net for mailing, but also for many other purposes; i.e. if the Web is becoming more and more integrated into the every-day activities of people, the readiness to participate in Web survey may increase.

Presently, there is still a strong selection bias caused by gender and by familiarity with the net. The participation rate in the subsequent Web survey was much higher for males than for females compared with the rates of Internet use in general. This means that the subsample of people participating in the Web survey does not reflect the structure of the subsample of Internet users in the CATI sample because of a systematic bias. If a panel study is intended, in addition, other systematic effects have to be accounted for. In the present study about 70% of those who participated in the Web survey agreed to participate once more.

One further problem remaining is the potential selection bias in the total CATI sample. Since in Germany refusers are not contacted again, we do not know much about them. The recruitment of participants in this study was done within a survey mainly addressed to a topic other than the address recruitment; i.e. recruitment was not the main purpose of the survey. As is well-known from the literature, the quality of the CATI sample depends on several factors, among these the attractivity of the topic. In the future, more investigations are needed to evaluate the effect of recruitment by a survey with sole screening purposes. A further important factor determining the participation rate for the Web survey is the purpose of the Web questionnaire. In the present case, the Web questionnaire had a close relationship to the preceding CATI interview and had a recognizable scientific purpose, but just this doesn’t hold in other domains like e.g. marketing research.

References

Biemer, P.P. & Trewin, D. (1997). A Review of Measurement Error Effects on the Analysis of Survey Data. In L. Lyberg et al. (eds.), *Survey Measurement and Process Quality*. New York: Wiley, pp. 603-632

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, pp. 464-494

Dillman, D. (1978). *Mail and Telephone Surveys: The Total Design Method.* New York: Wiley.

Dillman, D. (2000). *Mail and Internet Surveys: The Tailored Design Method.* New York: Wiley.

Gabler, S. & Häder, S. (1997). Überlegungen zu einem Stichprobendesign für Deutschland. *ZUMA-Nachrichten* 41, 7-18

Gabler, S. & Häder, S. (1998). Probleme bei der Anwendung von RLD-Verfahren. In S. Gabler, S. Häder &. J. Hoffmeyer-Zlotnik. (eds.), *Telefonstichproben in Deutschland.* Opladen: Westdeutscher Verlag, pp. 58-68

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. & Couper, M.P. (1998). *Non-response in Household Interview Surveys*. New York: Wiley.

Nicholls II, W.L., Baker, R.P. & Martin, J. (1997). The Effect of New Data Collection Technologies on Survey Data Quality. In L. Lyberg et al. (eds.), *Survey Measurement and Process Quality*. New York: Wiley, pp. 221-248.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

About the Authors

Marc Deutschmann is a Survey Methodologist at the Social Survey Research Centre of the University of Duisburg, Germany.

Frank Faulbaum is Professor of Social Research Methods and Statistics at the University of Duisburg, Germany and Director of the Social Survey Research Centre.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

Web versus Mail Questionnaire for an Institutional Survey

Katja Lozar Manfreda, Vasja Vehovar, Zenel Batagelj

Abstract

The Web survey mode often results in large non-coverage and non-response errors. However, in populations with high Internet penetration rates, this problem is much less severe. This is also the basic finding of this paper where three experimental groups were tested in a survey of school institutions. The response rates and the data quality of the Web survey were comparable to the pure mail option, where the costs were much higher. On the other hand, surprisingly, the mixed mode approach did not prove very helpful. This paper thus contributes to the understanding of the Web survey method (http://www.websm.org) showing that we are still searching for a proper non-response strategy for the Web survey mode.

Keywords

Web surveys, response rates, contact strategies, cost-benefit evaluation, mixed mode data collection, mode differences

1. Introduction

With Web surveys the probability samples of the general population generally suffer from large survey errors due to non-coverage. The persons not using the Internet are often excluded from these surveys and responding units are thus not representative for the population (Flemming and Sonner, 1999; Hoffman and Novak, 1998; Schmidt, 1997). In addition, response rates in Web surveys are also relatively low, so the validity of this survey mode is relatively questionable. However, much more encouraging news is arriving from establishment surveys (Clayton and Werking 1998).

Sometimes Web surveys are also used in a mixed mode design in order to overcome the problem of non-coverage. The Web can be used to survey units with Internet access, while more expensive methods can be used to survey those without access. In addition, mixed-mode strategy can be used also to overcome the problem of non-response, e.g. the Web survey can be used in the first wave starting point, and more expensive methods can be used to handle non-respondents.

This paper tested the alternative procedures for conducting Web surveys, using a mixed-mode strategy. The population of interest was represented by Slovenian primary and secondary schools. Previous research has shown that the majority of them already have access to the Internet (95% in 1999), therefore a Web survey seemed an attractive alternative.

Results from an experimental test of two procedures for conducting Web surveys, one pure Web survey, the other mixed-mode (Web and mail), are compared to a mail survey control group within the same population.

In the mixed-mode group, respondents were first invited to answer a Web questionnaire. In the reminder, they were invited to ask for a paper questionnaire, if they could not answer the Web questionnaire. In the second follow-up, all the remaining non-respondents in this group received also the paper questionnaire in addition to the invitation to answer the Web questionnaire. With this strategy we offered a chance to participate in the survey even if the intended responded did not have access to the Internet (therefore to overcome the problem of non-coverage). In addition, we maximised the effort to increase the response rate, while keeping the costs as low as possible. We begun with a Web survey as a cheaper alternative, and then used progressively more expensive methods for non-respondents. We assumed that additional methods would have a larger impact on the response since individuals may have a mode preference (Goyder, 1987; Groves and Kahn, 1979). Response rates and costs are therefore compared in the three experimental groups.

Of course, the overall quality of data should also be high in a Web survey mode, so we should get comparable substantive results and data quality. Previous comparisons of Web and mail surveys have given mixed results. For example, Gonier (1999) reports on several comparisons of marketing Web surveys with telephone, mail and mall surveys in the USA. He showed that a carefully designed Web survey could lead to same business decisions as other methods. Web respondents offer more extreme answers, however the means are comparable. Kwak and Radler (1999) report on a comparison of a Web and a mail survey of students. The response rate in the Web survey with e-mail invitation was lower (27%) than in the mail survey (42%). Comparison of respondents' answers showed similar results regarding the psychological measures, however differences in attitudes regarding technology, some technological characteristics and demography. In addition, there were also differences in relationships among the variables. Lozar Manfreda et al. (2000) compared results from a mail and a Web survey of business establishments with access to the Internet. While response was somewhat lower for the Web survey (26% in comparison to 39%), there were very few differences in the respondents' answers and some differences in the data quality.

Mixed results of previous studies suggest a further exploration of the problem of the possible mode effect in Web surveys. Therefore, not only response and costs, but also differences in substantive responses and data quality should be compared.

2. The study

Background

The project RIS – Research on Internet in Slovenia (http://www.ris.org), carried out at the Faculty of Social Sciences, University of Ljubljana, conducts a variety of regular surveys since 1996. The survey on the usage of Internet in educational institutions in December 2000 was thus performed for the fifth time since 1996. When speaking about educational institutions in Slovenia, it should be noted that there are 445 elementary schools and 155 secondary schools. All the units were included into the survey, so no sampling was applied.

This annual survey of educational institutions has always used a paper questionnaire administered by mail. In the 1998 survey, a Web questionnaire was also prepared and respondents could optionally choose, whether to answer the paper or the Web questionnaire. However, extremely low response was obtained through the Web at that time, although the Internet coverage itself was not a problem, but rather the low preference for the Web mode. This could be attributed to the specific efforts needed to answer such a questionnaire. While the paper questionnaire was already at hand, for the Web questionnaire the respondent needed to access the Internet. Similar results were reported also in another study of business establishments where respondents had a chance to choose between the paper and Web questionnaire (Vehovar et al., 2000b). There, only 15% of the final respondents used the Web option, the rest of them preferred the mail and a smaller part used the fax questionnaire.

Nevertheless, the complexity of the questionnaire (several skips and different questions for institutions already using the Internet and those planning to use it in the near future) and the resulting extensive rates of item non-response suggested, once more, the option of computer-assisted survey data collection for the 2000 study. The cost considerations are also advantageous for the Web survey mode. Due to the high Internet penetration rate (the RIS 1999 survey revealed that over 95% of primary and secondary schools have access to the Internet), a Web survey seemed an attractive alternative. In order to avoid the mistake from 1998, another survey design was planned, as explained further.

Methodology

Institutions (secondary and primary schools) were randomly assigned to three experimental groups, each of size n=200.

1. **Mail group**: Traditional mail survey with two-follow-ups: a reminder and a letter with a replacement questionnaire (TDM procedure);
2. **Web-mail group**: Mail advance letter with an invitation to a Web questionnaire, a reminder with an invitation to ask for a paper questionnaire if response through the Internet is not possible; a follow-up letter with the paper questionnaire for all remaining non-respondents;
3. **Web group**: Mail advance letter with an invitation to a Web questionnaire only and two similar mail follow-up letters.

No incentives were used. The paper and the Web questionnaire were similar in the sense of the same questions and answers and same question order. In order to resemble the paper questionnaire the design of the Web questionnaire was very simple. There were several questions on one HTML page and a new page appeared only due to automated skips.

An ID number was printed in the lower right corner of the paper questionnaire. For the Web questionnaire, respondents needed to enter their ID number at the survey introductory Web site in order to access it. The schools' principal or person responsible for the computer technology at the school was asked to answer the questionnaire.

From previous surveys we knew that over 95% of schools have access to the Internet, therefore no special instructions were given in case a school would not have access to the Internet. We assumed that in the third group – only the Web questionnaire – the respondents would ignore the quest for participation or try to answer the questions from another location, not from the school, in the event that the school did not have access to the Internet. It turned out that this was not a problem at all. Every school that participated in the survey from the first experimental group – mail survey – had access to the Internet. In the other two groups 4 Web respondents answered that they still did not have access to the Internet, but they planned it. Nevertheless, they obviously had a possibility to answer the questionnaire from another location, where access to the Internet was available. Therefore we can assume that a large majority of units in our sample have access to the Internet and no special caution in comparing response from different experimental groups due to the possibility of no Internet access is needed.

3. Results

Response rates

The highest response was achieved in the group in which the traditional mail survey was implemented: 89% of questionnaires were returned answered. For the Web survey, the response rate was 77% which is also very high in comparison to the response rates in Web surveys previously published in literature. The option to answer the paper questionnaire in the second and the third follow-up actually increased response for a few percentages, up to 80.5%, however the increase is not statistically significant. Only a few participants asked for the paper questionnaire after the first follow-up (8 or 5% of non-respondents after the 1st mailing), however, most non-respondents decided for the paper questionnaire when so offered in the second follow-up (27 or 34% of non-respondents after the 1st reminder).

Table 1: Response for experimental groups

		1st contact	1st follow-up	2nd follow-up	Total	
Mail		57.0%	18.5%	13.5%		89.0%
Web+mail	Total	21.4%	38.7%	20.4%		80.5%
	Web	21.4%	34.7%	6.9%	63.0%	
	paper	/	4.0%	13.5%	17.5%	
Web		27.7%	33.1%	16.2%		77.0%

The difference in response between the mail and the other two groups was the largest after the first contact. The reminder significantly increased the response rate for the two groups with the Web questionnaire. Similar findings on the importance of the follow-ups in institutional Web surveys were found also by Vehovar et al. (2000b). The effect of follow-ups was larger for the group where also a paper questionnaire was offered in addition to the Web questionnaire.

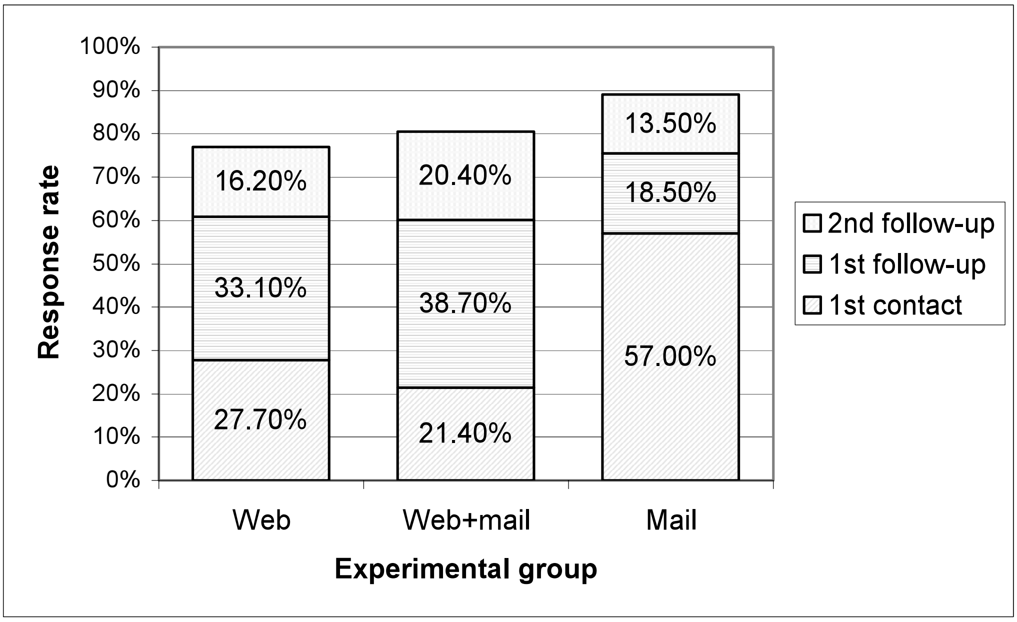
Despite the fact that the response for the Web questionnaire was somewhat lower in comparison to the paper one, it was still high in comparison to other published studies on Web surveys. We can contribute this to the traditional high response in this particular research project. Educational institutions are invited to participate in this survey every year since 1996 and one of the survey sponsors is also the Ministry of Education. In addition, two follow-ups are used, help from our office is offered on weekdays and the questionnaire is made as simple as possible. Beside that, information-communication technology is known well enough within the schools in order for them to be able to answer a Web questionnaire.

Costs

When calculating costs for the implementation of the survey in the three experimental groups we included only variable costs. Costs for developing a Web or a paper questionnaire and computer-assisted form for data entry from the paper questionnaire were not included since they were the same in all cases. Namely, we used the process of Integrated Computer-Assisted Data Collection (ICDC) - the software that enables easy transformation of a survey questionnaire to different survey modes (Vehovar and Batagelj, 1996). We also did not include additional costs for adequate software and hardware equipment.

The calculations therefore included costs for envelopes, stamps, advance letters, paper questionnaires, data entry for the paper questionnaire, and accompanying administrative work, as needed for the implementation of the survey in each experimental group. As expected, the costs were the largest (320 EUR) for the pure mail group due to additional costs for questionnaires, envelopes and stamps for re-

Figure 1: Response rates for three experimental groups



turning questionnaires, and data entry. The costs for the pure Web group were about half of the mail group (165 EUR).

For the mixed-mode group the costs amounted to 210 EUR. For the increase in the response rate from 77% to 80.5% the costs increased for almost one third. For the same budged as used for the mixed-mode group with 200 units in the initial sample, a pure Web survey with an initial sample of 255 units could be conducted, which would result in smaller sample variance. Or, additional funds could be allocated to some aspects of data quality, such as further questionnaire testing.

Differences in responses

In addition to the achieved response rate and related costs, the decision about choosing the survey mode is based also on potential differences in responses. We therefore compared the responses from the three groups. An analysis of variance was used for comparing the mean score on interval and scale variables and percentage of interest on nominal variables. The Bonferroni or Games-Howell post-hoc tests were used for testing the differences in pairs of groups.

Substantial differences

For most of the items there was no difference in the respondents' answers. Statistically significant difference (at $\alpha < 0.05$) occurred only at 8% of items (11 our of 138). There is no pattern where the difference occurred. For example, respondents from the mail group have access to the Internet somewhat less time than respondents from the other two groups. Also, respondents from the mail group claim the Internet to be somewhat less important for their institution as respondents from the pure Web group. Respondents from the mail-Web group showed the least agreement with the statement regarding the positive role of the government in the Internet development within the education system and this is statistically significant from the Web group which showed the highest agreement with this statement. The mail group is somewhere in between. There was also difference in the frequency of Internet use at certain classes or for certain purposes, however there was no pattern in the responses.

Item non-response for closed questions

For each respondent the percentage of items where non-response or "don't know" were given were calculated. The average item non-response for the group where only the paper questionnaire was used was 9%. The average item non-response for the group where the Web questionnaire was used was 23% and the difference is statistically significant at $\alpha < 0.01$. The item non-response for the group where the paper questionnaire was offered in the follow-ups was 20%, which is statistically significantly different from the group with only the paper questionnaire, but not from the group with only the Web questionnaire.

The time when the questionnaire was answered – after which contact (1st mailing or after follow-ups) did not significantly influence the item non-response.

The item non-response is definitely larger for the Web questionnaire. Even if we exclude partial responses from the Web questionnaires (7 respondents who only answered questions on the first few screens, but then quitted) the differences are still significant (9% in the mail group vs. 19% in the combined Web-mail group vs. 21% in the pure Web group).

This is consistent with the finding from another establishment survey, a survey of business companies (Lozar Manfreda et al., 2000). In both cases, no forced reminders for the item non-response were used.

Without this programming feature, it looks like the item non-response is larger in Web questionnaires than in paper questionnaires. In our case reminders were not used since only higher versions of browsers would support them (those that support Java Scripts) and we were not sure whether our respondents use them.

Item non-response for open-ended questions

In 13 cases the questions were open-ended asking for a specific number, e.g. the number of employees, of pupils, of employees with access to WWW within school, of computers with certain characteristic, etc. Percentage of these items that were not responded or the response did not prove to be useful (such as 'almost all' instead of the exact number) was not statistically significantly different within the three groups (8% in the pure mail group, 11% in the combined Web-mail group and 10% in the pure Web group).

Respondents were also invited to give comments at the end of the questionnaire. Comments were given somewhat more often from respondents in the pure mail group (11%) than in the other two groups (9% in the combined Web-mail and 8% in the pure Web group), however the difference is not statistically significant.

Respondents' satisfaction

At the end of the questionnaire we asked respondents to evaluate the survey according to several criteria (subject, difficulty, length, graphics) on a scale from 1 (not adequate) to 5 (excellent). Average score on these items was the same in all the groups, expect for the graphical layout of the questionnaire. Web respondents liked the layout of the questionnaire more than mail respondents. In addition, respondents from the Web and combined Web-mail groups also evaluated the survey overall somewhat higher.

4. Conclusions

In this paper we described an experiment where the Web survey mode was compared to a mail survey and also to a mixed-mode survey design. We showed the following:

- The Web survey gains only slightly lower response compared to the mixed mode option. The mixed mode option was thus not really fruitful for dealing with non-coverage or non-response. Compared to the pure Web option the results were almost the same, but costs and administrative inconveniences were much higher. Together with a previous discouraging experience where the units themselves have the option to select among the modes – there, again, were no benefits but only complications in survey administration - the mixed mode option does not seem very advantageous.
- The difference to the mail survey was also relatively small. The mail solicitation thus provided sufficient response rate also for the Web survey mode (77%). There was also not much difference in substantive results among the three experimental groups, suggesting that despite the slightly lower response rate in the Web survey, the Web survey presents an attractive alternative to the mail survey since the costs are dramatically lower.
- The Web survey definitely needs at least two reminders when solicitation is performed with mail. The follow-ups are much more fruitful compared to the pure mail option.

- It is a little worrying that the item non-response was much larger for the Web questionnaire when no forced reminders were applied. It seems as if people are less committed to completely answer the Web than the paper questionnaire, unless reminders for item non-response force them to answer; this may also have other disadvantages.

A general solution is thus still needed to obtain effective co-operation in Web surveys, an approach similar to the Total Design Method (Dillman, 1978) for mail surveys. The Tailored Design Method (Dillman, 2000) presents an initial step in this direction. As long as we miss an optimal strategy also the comparisons with other modes – that were already explored – could not be completely fair.

We can conclude that in this survey the Web questionnaire can be used in future instead of the mail survey, since it preserves the quality and remarkably decreases the survey costs. Additional funds could be allocated in further testing in order to decrease the item non-response.

In future, the experiments should be directed to forced reminders, because it is possible, that this tool may increase partial non-response and also decrease the satisfaction with the filling of the questionnaire.

References

Dillman, Don A. (1978): Mail and Telephone Surveys. The Total Design Method. John Wiley & Sons, New York.

Dillman, Don A. (2000): Mail and Internet Surveys. The Tailored Design Method. Wiley-Interscience. John Wiley Company: New York, NY.

Flemming, Greg, and Sonner, Molly (1999): "Can Internet Polling Work? Strategies for Conducting Public Opinion Surveys Online". Paper presented at the 1999 AAPOR Conference, St. Petersburg, Florida, May 13-16, 1999.

Gonier, Dennis E. (1999): "The Emperor Gets New Clothes". ARF's Online Research Day - Towards Validation. Advertising Research Foundation, New York, pp. 8-13.

Goyder, John C. (1987): The Silent Minority: Non-respondents on Sample Surveys. Westview Press, Boulder, CO.

Groves, Robert M., and Kahn, Robert (1979): Survey Research by Telephone. Academic Press, New York.

Hoffman, Donna L. and Novak, Thomas P. (1998): "Bridging the Racial Divide on the Internet". Science, 280, 390-391.

Kwak, Nojin and Radler, Barry T. (1999): "A Comparison between Mail and Web-based Surveys: Response Pattern, Data Quality, and Characteristics of Respondents". Paper presented at 1999 Annual Research Conference, organised by Midwest Association for Public Opinion Research, November 19-20, 1999, Chicago, Illinois, USA.

Lozar Manfreda, Katja, Vehovar, Vasja, and Batagelj, Zenel (2000): "Veljavnost interenta kot anketnega orodja". Teorija in praksa, 37, 6, pp. 1035-1051. (Engl. Validity of Internet as a survey method)

Schmidt, William C. (1997): "World-Wide Web survey research: Benefits, potential problems, and solutions". Behaviour Research Methods, Instruments, & Computers, 29, 2, pp. 274-279.

Vehovar, Vasja, and Batagelj, Zenel (1996): "The Methodological Issues in WWW Surveys". Paper presented at CASIC '96, San Antonio. Available: http://www.ris.org/casic96/.

Vehovar, Vasja, Lozar Manfreda, Katja, and Batagelj, Zenel (2000a): "Participation in solicited Web surveys: Who comes farthest?" Paper presented at Fifth International Conference on Social Science Methodology, Cologne, Germany, October 3-6, 2000.

Vehovar, Vasja, Lozar Manfreda, Katja, and Batagelj, Zenel (2000b): "Sensitivity of E-Commerce measurement to survey instrument". In: KLEIN, Stefan, O'KEEFE, Bob, GRIČAR, Jože, PODLOGAR, Mateja (Eds.). Thirteenth Bled Electronic Commerce Conference, Bled, Slovenia, June 19-21, 2000. The end of the beginning : proceedings. Kranj: Moderna organizacija, 2000, pp. 528-543.

About the Authors

Katja Lozar Manfreda, M.Sc., (katja.lozar@uni-lj.si) is a Ph.D. student at the Faculty of Social Sciences, University of Ljubljana. Her research interests include methodology of Web surveys, and particularly the non-response in Web surveys. She is the author of the leading Web site on Web survey methodology (http://www.websm.org).

Dr. Vasja Vehovar (vasja.vehovar@uni-lj.si) is a professor at the Faculty of Social Sciences, University of Ljubljana. He was educated at the University of Ljubljana (PhD), the University of Essex (MA), and he was a Fullbright scholar at the University of Michigan, Institute for Social Research. His research interests include survey methodology, particularly the survey sampling and Web survey methodology. On the other hand he also conducts substantial research on the Internet and e-commerce. Since 1996 he has been the principal investigator of the national research on Internet in Slovenia (http://www.ris.org). He published in the Journal of American Statistical Association, Journal of Official Statistics, as well as chapters in the books of major publishers (Wiley, Greenwood Publishing Group).

Zenel Batagelj (zby@cati.si) is a Ph.D. student at the Faculty of Social Sciences, University of Ljubljana, and research director at a private research company CATI Center, which is the leading market research company for telephone and Web surveys in Slovenia and also the partner of Gallup International. He developed his own software for the integrated computer support for all modes of surveying. His research interests include methodology of Web surveys, particularly the design issues, as well as the survey measurement of information-communication technologies.

Faculty of Social Sciences, University of Ljubljana
Kardeljeva ploščad 5, 1000 Ljubljana, Slovenia
Tel: + 386 1 5805 100, Fax: +386 1 5805 101

CATI Center
Tržaška 2, 1000 Ljubljana, Slovenia
Tel: + 386 1 4211 910, Fax: +386 1 4211 970

Dissemination of Statistical Information

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

Seizing Opportunities to Inform the Information Society: creating Value from Statistics through exploiting New Technologies

Len Cook

Keywords

Official statistics, National Statistics website, Internet, Information Society

1. Introduction

Statistical data is very important to public policy, and to public understanding of the context of public policy. Its accessibility contributes to the effectiveness and trust in government, and the efficiency and stability of financial, commodity, labour and other markets. The effective delivery of official statistics on the internet is not only driven by the potential efficiencies for continuing to deliver existing services to existing users, but the growing expectation of access at a low cost by citizens to government services, higher levels of numeracy and a sea change in the ways in which statistics are able to be made interesting and relevant. Our future service model may well come from the media, compared with the present more academic-like approach we take.

The internet is a fast growing, highly integrated global facility, and the competitive position of the statistical office internet service, its comparative qualities, and the nature of relevant innovations, are highly visible through their being provided alongside other public services, and alongside those of all other information providers, in each and any of the situations users are in when they draw on our services.

2. The special value of the Internet to official statistics

The internet is an architecturally independent pathway that enables myriads of linkages through public and private communications networks between a large minority of private households, many businesses, most public bodies, many community organisations and nearly all of the media, research and international organisations. The internet enables a gigantic leap in the numbers of people who can access official statistics at a very low cost, regardless of region, physical mobility or language.

The internet has particular value to official statistics, because of the capacity to present statistical information dynamically, in visual form through graphs and video, through real-time linkages to databases, and linked to contextual information. It provides for a major step change from the static nature of books, whether in printed form or Acrobat file. The ability to make statistics understandable, interesting and relevant can increase greatly, at little cost.

Through the power of contemporary web based tools in providing content sourced from databases, the internet enables powerful searching of information with or without structure, making accessible huge repositories of information perhaps originally organised in ways that were only relevant to the uses that the information was initially designed to be used for. Users can easily gain familiarity with what is available, and readily retrieve it, and learn of whether similar material exists, with little time or skill. The users' desktops can be customised to fit their interests, in the mapping, searching and presenting of information.

The internet can enable information to be efficiently provided in the great variety of forms relevant to official statistics, particularly text, statistical tables, graphs, maps. Such releases can be made available in complete studies, as tightly focused pictures of the state, organised to whatever extent is necessary to make a reasonably complete parcel of information.

The internet can integrate data, voice, visual and other means of interaction. This enables each and every communication to be concluded in a satisfying way, across the huge variety of circumstances through which users of statistics engage with official statistics, regardless of how it was initiated. It provides for the resulting service delivered to users to be able to be customised to their needs and ability to absorb statistical results.

Monitoring of activities of the statistical office web site, including degree and form of access, enables patterns of behaviour, levels of use, and reactions to change to be cheaply and quickly assessed, at the level of the system, the specific service, and the individual user. Changes in demand and the value placed on new services can be quickly and accurately measured, in aggregate, and for individual or groups of users.

The internet enables a high level of integration with related services, including billing and promotion of new material. It also facilitates the efficient, downstream production of internet based services by added value information providers

3. The nature of the statistical office web-site

The statistical office web site has the potential, among other things to become the prime means of delivering official statistics, managing statistical enquiries and meeting obligations to provide access to information about statistical practice. All statistical office services could be designed to service web based access, and other forms of access will be efficient by-products of the web service and associated facilities, with few exceptions.

The statistical office web site could be a wholly integrated part of internal statistical office statistical and business systems. Information will be organised so that the power of the web tools and browsers for unstructured searches can compliment any structured access paths. The web focus of all internal processes will not usually need any intervention in business processes to provide for the results to be delivered on the web, and any that is required will be, of necessity, quite limited.

Statistical material and other information and web site management will be through presentations that are designed to appear firstly as quality presentations on an internet screen. Where printed, or where delivered in machine-readable form, such as Excel spreadsheets, that will determine the design of the information.

A typical web site might consist of five quite distinct elements, which will be linked where relevant.

Pictures of the nation: These will be a series of graphical and text based presentations with dynamic elements, which present self contained and usually short stories about particular topics. These will be dynamically updated, where they are focused on the latest available measures. They may contain videos, audio, graphics and other tools embedded in them, that provide a dynamic and interactive element. There will be a high editorial contribution from the statistical office, including design and selection not only of measures but how they are presented. Particular elements will be:

- the analysis and main results associated with each first release of statistics
- the nation's yearbook in a new form
- topic reports that reflect the evolution of existing services such economic trends, social trends, labour trends
- special analytical reports
- high profile statistics
- major new statistics (e.g. Census 2001)

Information gathering: This will provide structured and unstructured access to:

- the formal publications and presentations of the statistical office, as well as available but unpublished tabulations and series.
- the structured statistical data-banks necessary for making available at low cost repeated access to indexed, machine readable data, particularly time series, regional statistical aggregates and accompanying aerial descriptions, and multi-way cross tabulations, most of which are based on a few critical cross cutting variables such as family, industry, sector, commodity, population group, age or sex.
- statistical summary data-files in special areas particularly foreign trade, where a well defined set of analyses of individual import or export commodities exists.

Secure access to confidential unit record statistical data sources: Retrieving information from survey and administrative data sources:

- statistical unit record database extractions, where users can define tables or other aggregates to be extracted from unit record databases using flexible self specified enquiries, and receiving results that are tested against confidentiality protection rules.
- modelling environments, which may be made up of a mix of unit records, statistical summary information or other statistics, and to which particular models can be applied, or which are linked to generalised modelling environments such as SAS.

Documentation: This will provide structured and unstructured access to:

- writing about statistical practice, statistical processes, methods and tools. It will also provide access to meta-data of all forms, including classifications, definitions, frameworks such as ESA,
- information about each statistical survey
- information about the statistical office
- information about National Statistics

- catalogue of services
- National Statistics strategies and work programme
- links to international and other relevant country web-sites.

Special services: These will be special services that develop for particular user communities, focus on particular products, or increase visibility of something important. Examples are:

- education services
- communications channels (feedback, email, voice messages, trigger for personal contact)
- media services

4. Distinct faces of the statistical office web-site

The statistical office web site may need to have two quite distinct faces, each of which will present a distinct menu of the available statistical office internet services.

1. *A public face,* which will be focused on immediately usable material, free access, supported by a fixed amount of telephone support each year. All documentation will be available, and several information gathering services will be available even where chargeable, so that users can learn of their existence. The public face will be designed to need no training for its use. The monitoring of each user's access will enable some customising of feedback to users, and we could plan to have a variety of front ends to the web site, if distinct groups of users had specific and particularly different needs that justified that. There will be an integration of voice with data.
2. *A "subscriber managed" face*, that will provide services that are either needing access to confidential records held only in statistical office custody, or which use services that are provided on a subscriber basis, or which need training and support for their effective use, such as the time series access service. The needs of each and every user of this service will be of sufficient value to the statistical office, or of sufficient importance to government, that they might justify their own access path through the web-site services, as well as their own approach to linking to parts of the web-site content either on their own internal network, through the internet or through some private network. The integration of the statistical office web site with other services used by these users could be customised to their needs. Training will be available for services that need it. Security will be comprehensive, and user specific, using civil service wide infrastructures where available. There will be an integration of voice, data and video. A chargeable service enabling access to statistical office experts will be available.

Having two distinct web faces will ensure that the public face is consistent with free access to statistics, has the highest level of consistency in the look and feel of its services, has minimal security overhead, and is reproducible in many ways. The subscriber managed face will firstly be designed to deliver to policy ministries a comprehensive, easy to use, always available and well supported access to statistical information to ensure that all relevant information is discovered and accessible, for whatever question is being addressed. Other specialist users outside government will also use these specialist services, which will be more limited for unit record information except where there is the same legal authority to have access in the form potentially available.

5. Integration with production processes

The statistical office web site will be fully integrated with statistical office systems and processes. Document management in some sites such as Statistics New Zealand is managed through mirroring the information provided in a Lotus Notes based knowledge management system, so that the marginal costs of transferring information into the web environment includes no translation cost. The preparation of statistical pictures should involve tools that belong to the statistical office standard tool-kit, and which by design integrate at no extra cost into the web environment. This might include Microsoft tools such as Excel, or necessitate a new standard graphing tool. The dynamic access to statistical measures needs contemporary web integratable summary data management tools for time series (Fame is a good example), regional data (Supermap is a good example), and multi-way tables (Supertable or IBM Pivot table are good examples). Statistical office tools in this area may well all need upgrading.

The information management capability should accept enquiries on statistical records driven by web managed tools, and should provide extractions of data in a form that readily inputs into the standard web integrated summary data management tools.

There should be no web content that requires operator intervention or special processes to translate the information onto the web site. The consequence of this is that the statistical office web site will be developed in parallel with information management initiatives.

6. Web based improvements in the presentation of official statistics.

The web creates the potential for the presentation of official statistics to radically improve in quality in several ways. Each of these will challenge the existing practices of the statistical office, but also will generate more added value from existing major work such as the studies that make up the UK Social Trends. We will see:

- timeliness improvements
- quick extraction by users of indexed, machine readable statistical series, and of graphics.
- greater emphasis on graphics, and maps
- new forms of presentation
- more relevant integration of information
- dynamic response when changing the cross cutting variables of any analysis
- ease of quick linkage to related sources of information, and to relevant statistical services
- easy links to meta data associated with any series
- easy fit of results of enquiries with other electronically provided outputs

7. Implications of a centralised web publishing service

The statistical office information technology architecture will embrace the internet, creating a wholly integrated environment. This integration covers technology, security and information, involves tools, systems, processes, management, client, suppliers, reporting. We will create high value from communicating via databases, and have efficiencies gained by structuring information that was previously unstructured. There will be a very large web site with authorship decentralised but control centralised. There will be considerable ease of exchange of information with partners who use Lotus Notes. The statistical process and data access benefits we will obtain from the fully web integrated architecture will be:

- extensive public documentation at low cost
- hugely improved performance monitoring
- lower cost structure for all activities
- reduced cycle times for new activity
- credible capacity for more innovation
- more rapid creation of new products
- able to manage more change at any time
- Web can provide several channels for supply and delivery of services
- much reduced communication costs
- easy links to any relevant organisation
- all statistical office statistical services are able to be digitised
- Web compatible services penetrating down into all business processes
- exchangeable business systems and processes

The new statistical office web site will be an statistical office wide facility, relevant to all statistical outputs and services, and all statistical outputs and services will use it, as the prime means of service delivery. This will require strong corporate standards that are relevant to all the work of the statistical office, state of the art corporate systems that are used locally, everywhere, and the application of corporate practices at all times. To facilitate this, the integration of technology environments will necessitate a high degree of standardisation of systems and equipment, integrated software and data management tools, one statistical office publishing policy, and one release process. We need to determine our corporate tools, e.g. Lotus Notes, and apply universally without question. A strategy for settling on new publishing approaches for statistics needs to be prepared. We need to decide on what sort of books and other printed material we will continue with, to balance what the web can now best deliver.

The new statistical office web site will be tightly integrated with the new statistical office information management and knowledge management tools, and the developments will be planned and implemented in parallel.

8. So what is ONS doing to respond to these imperatives?

The challenges described above amount to no less than a revolution in the way statistical offices hold, manage, give access to and think about information. In ONS we have recognised that the management and culture changes required are fundamental will require sustained effort over many years.

We have instituted a comprehensive business strategy with better management of information and stronger statistical infrastructure as the centrepiece supported by reformed business processes and rigorous change management - covering both the harder aspects of business planning (including project and financial management and office systems) and the softer side of culture change and development of the skills and confidence of our staff to thrive in the new and ever changing world.

This strategy will take time to deliver results and may be seen as having a five-year time horizon. The world will not stand still and wait for us, however. To meet the need for us to deliver performance gains in the shorter term we have initiated an Information Age Access Programme which will establish a new website for National Statistics by January 2002. This is a tactical programme designed to liberate the information resources we already have and make them easily accessible and available over the web.

The programme comprises six distinct elements. The aim is to put the web at the heart of how we communicate statistics. The centrepiece is the new website itself which will aim to provide all the information resources of National Statistics as though they were one single publication describing the statistical state of the nation.

The second element is a range of access services allowing mass customisation of access reflecting the full range of needs. These range from the novice visitor to the site who wishes to browse or search simply and quickly for information with no prior knowledge of ONS or its processes to the regular user who uses large volumes of information to analyse in depth and simply needs to get desktop access to the detail fast and reliably. Access services might also include integration of telephone enquiries and opportunities to make use of new media to communicate statistics based on the web offering e.g. via WAP phones.

These access services also include, of course, a range of paper publishing options for those many users of statistics who wish and are likely to continue to wish to have paper documents. The difference will be that the paper copies will be created from the common electronic information pool rather than vice versa.

The third element is an internal interface system. A structured basis for managing the content of the site and ensuring that the right people create the right material at the right time to the right set of corporate standards is being created. This will feed information into a single publishing repository allowing ultimate delivery to be from any one of a number of publishing media. It is in this element that we will begin to effect the cultural shift needed across the business. Consequently this element will include a major emphasis on awareness raising and training across the Office and the wider National Statistics community.

These three elements form the core of the programme but are complemented by three further projects which will enable maximum leverage to be gained for statistics users and will help position ONS as a leader in e-government. The first is simply to keep making improvements to the existing website between now and January 2002: we can never stand still. The second is to join up the ONS web offering with the cross-government Knowledge Network, providing ready access to important facts and figures on-line. The third, and more wide-ranging still, seeking to position ONS as a leader in e-government playing a major role in a range of wider initiatives.

9. Conclusions

For any statistical office, the web site is a critical initiative. We need some clear direction on its focus, integration with statistical office infrastructure, and the potential changes that it will make possible. The technology options are considerable, and when we have succeeded in them, we will find that business change may be slower than that of the technology. New technologies that exist now give statistical offices the chance to be at the forefront of public sector modernisation, provided we rethink; and redevelop the statistical services we provide. The internet will be fully exploited by the national statistical office when there is a civil service wide secure network to work within, all agencies have a similar level of competence in IT matters, and public access to the internet is judged equitable. This paper is about what we can achieve with or without these broader successes, and about the contribution we can make to making these successes visible when they occur.

About the Author

Len Cook is the National Statistician and Registrar General for England and Wales. He joined ONS in May 2000 having previously held the post of New Zealand Statistician since 1992. He has a particular interest in social policy, superannuation, taxation, demography, statistical methodology, marketing, and the application of information technology to information issues. In addition to other interests, he has been actively involved in the presentation of statistical trends to public and professional audiences.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

Publishing a Major Government Report in Electronic Format - the Implications for Social Researchers

Alison Walker and Steven Connor

Abstract

In line with government guidelines, ONS is pursuing a policy of web-based dissemination. The 'Living in Britain 2000' report was selected as a pilot project with the objective of publishing a major report in electronic format, using a variety of web-based technologies, to enhance both the customer usability of the publication and the efficiency of the ONS production process.

Living In Britain presents the results from the General Household Survey, a multi-purpose survey that runs continuously and provides results on an annual basis. The 2000/1 edition in printed format is due for publication in November 2001.

The paper will describe the development of this 'write for multi-media' project and discuss the consequences of these changes from the viewpoint of the social researchers involved in production of the report.

Keywords

Multi media publishing, content management, General Household Survey, Living in Britain, government social surveys.

1. Introduction

This paper looks at the requirements for the production of a web-based annual report of the results of a large government social survey and then describes the solutions in the more general context of multi media publication for the Office for National Statistics (ONS).

2. Background

Until recently, the statistical dissemination policy of the Office for National Statistics has focused mainly on hardcopy publications. Once statistics have been collected and collated, and once any associated documents have been finalised, the primary means for their dissemination has been through either press releases or paper publications of one sort or another. In some cases, data or reports have also been published using disks or CDs but these have tended to be produced as electronic replicas of hardcopy publications. This kind of publication system has been characterised as being the product of a 'Push', or 'Broadcast' culture.

Initially, the advent of the web and the development of official statistical websites had done little to change this culture. The kind of data made available via websites tended to mirror the kind of tables that appeared in releases and books. The concept of electronic publishing was largely confined to the procedure whereby outputs were placed on websites using a pdf format that exactly replicated the hardcopy versions of those outputs. Within ONS, much of this is due to the fact that many of ONS's 'repository' databases were originally built in order to facilitate the production of hardcopy outputs, and because, with a few exceptions, these databases had not been directly linked to the website. This led to the situation described as 'reverse engineering' whereby electronic data (within operational systems) are transformed into paper data (within publications) and then re-converted back into electronic data (within websites).

In order to move forward to utilise the new technologies available, ONS commissioned a consultancy with experience of both paper and electronic publications to identify feasible approaches for future production and dissemination, while meeting user needs for the right information through the chosen channel in the appropriate context. A report was presented in June 2000.

Three possible types of publishing were identified, which follow on from each other chronologically:

- broadcast – a bulk-product is produced (eg an annual report of a survey) which is tailored for the main customers, while other users have to accept what has been presented.
- navigational - this is the production of fit-for-purpose web products, based on enriched content, rich metadata and links between commentary and data, where the value for the user is in the aggregation of these products, including external sources. Since the full range of content and metadata is available users can tailor what they access to their own requirements. Thus a 'publication' becomes indistinguishable from a web site.
- interactive - where the interaction is between the user and provider, or between users and other users, and where the value of the information is a function of the number of people using it.

The broadcast approach described the situation most often found at ONS where a main paper publication was produced. With the move towards electronic dissemination, the trend has been to put the paper products in their entirety on the web, or extracts with links to data.

To move to the navigational method, a data repository is required, where everything (data, metadata, analysis, etc) is held electronically and which then becomes the source for all other outputs. The interfaces between the systems that draw on the data repository will need to be well defined and well supported. Such an approach could not be achieved immediately for the multitude of ONS outputs. Once a navigational method has been adopted it is then easier to move to an interactive approach that would allow us to develop the most responsive relationship with users of the data.

Following discussion of this report, Len Cook, the National Statistician, issued a paper 'Publishing UK National Statistics on the internet' stating that ONS should move towards putting all its publications on the web. Some publications should become digital almost immediately and all should be available in this form within two years. This view reflects the Government's target that all public services should be on-line by 2005.

One of the key conclusions of the feasibility work was that pilot projects should be set up to encourage buy-in, highlight problems and discover what was needed more generally. Three pilot projects were agreed based on the following types of publications:

- database publications - which are low volume publications mainly produced annually with little or no text
- first releases – which are regular, frequent concise publications which contain mainly figures
- 'Living in Britain' - the report of the General Household Survey.

3. The General Household Survey

The General Household Survey (GHS) is a multipurpose continuous survey conducted on an annual basis, which collects information on a range of topics from people living in private households in Great Britain. It includes a range of topics such as housing, employment, health and health-related behaviours, such as smoking and drinking, family history, qualifications, income and ownership of consumer durables and vehicles.

The results of the survey are published annually in a report called 'Living in Britain'.

This was seen as an ideal publication to choose as a pilot since both the content of the report and the variety of users reflected the range of issues we wanted to investigate. Although the report is an annual publication in a standard format there is some variation in content from year to year. It contains a balance of text, tables and charts. The majority of tables are reproduced identically each year but there is some variation, as topics are examined in different ways or new topics are introduced. Many tables present trends within each annual report but others can only be seen as a time series in the context of reports from previous years. In addition to the results, the report also contains a small proportion of the available metadata for the survey.

The general household survey is an important source of data for a very wide variety of users. These include:

- policy makers in central government – the survey is mainly sponsored by ONS but ten other Government departments also contribute to the cost. In particular the Department of Health and the Department of Environment, Transport and the Regions make a substantial contribution. These users require the data for various policy initiatives both as a source of estimates for which it is the best source and, more particularly, for analytic purposes to inform the development of new policies.
- academics – the survey provides a rich data source for research with its particular importance as a source of cross-topic analysis since we collect such a wide range of data that can be analysed at individual, family or household level.
- local government – comparisons can be made between data that have been collected at local level with those we present at national and regional level.
- students – the data sets and reports are used for teaching purposes and as source material for specific topics.
- commercial enterprise – varying from pension providers to manufacturers of contraceptives.

- the general public.

With such a wide variety of users the requirements placed on dissemination are extensive.

4. Current methods of dissemination

Currently, there are several methods of dissemination.

- the ‘Living in Britain’ report as already described
- data sets - sent to the clients and deposited at the UK Data Archive at Essex University along with all the necessary metadata
- sets of additional tables - provided for clients to accompany each Living in Britain report
- an information service where users of all types can obtain details about current or past data or metadata, copies of sections of the report and ad hoc tabulations. This service is extremely well used.

The 1998 report was the first ‘Living in Britain’ to be put on the national statistics website as a pdf file. This has proved useful for answering general enquiries – saving staff time and photocopying costs.

While there is a complete spectrum of users, the standard methods of dissemination are more polarised. On the one hand, there is the broad brush approach of the Living in Britain report which provides much of the policy users’ initial needs but is too encompassing (and expensive) for the majority of general users. On the other hand, there is the micro-data, eminently suitable for the academic researcher but too detailed and complex to access for more general data requirements.

A web-based publication (or ‘Living in Britain web site’) with its extensive possibilities of access is the obvious solution.

5. What do users require

In 1997 an extensive review of the General Household Survey was conducted which included looking at user requirements. More recently, discussions relating specifically to Living in Britain as a web-based product have taken place with the main GHS clients.

As one would expect, the main user requirements varied to a considerable extent. There was general consensus that a web publication should enable users to search, download and print off copies of text, tables or charts. Some users still wanted the paper publication with which they were familiar and which they felt gave ‘non-technological’ access to data they knew was there. Others required an electronic replica of the paper report that they could use as a reference point for other users – ie a pdf file. This would also satisfy the requirement of being able to reproduce exactly what is shown on the screen.

In terms of requirements for the web product itself, these were also many and varied. There was consistent agreement that there should be links between tables, text and charts so that, for example, by selecting a table reference the table would be shown. Different users wanted the data and text presented in different ways so the design had to incorporate the facility to choose whether to view tables with the interpretative commentary, view tables separately or commentary separately. Many users felt easy access to explanations of terms used in tables or the commentary would be helpful which meant putting links to a glossary. A straightforward index with links was also seen as necessary.

In terms of metadata, the concurrent preparation of an electronic user guide for the GHS has made it possible to add links providing more detail on the collection and processing of the presented figures, for example table headings could be linked to the relevant part of the questionnaire and/or to the specification of derived variables. We also needed to consider the possibility of putting in links to other data sources or to the source of the GHS micro data – the UK Data Archive at Essex University. Perhaps the most demanding requirement is to provide links to past data to show trend data for tables that in the paper report only show current year figures. This requires past data to be added to the web site, which could be quite a task in itself.

From the production point of view we also wanted to be able to view the product in proof-form and circulate it to clients for comment and amendments before it was generally available.

Of course, an over-riding requirement was that the report was available within the agreed timetable.

Faced with a wide range of possibilities of what could be included in a web-based publication and a limited timescale, the task was to prioritise. However, a major advantage to both the producer and user is the flexibility of using electronic dissemination; the product is not a 'one-off' production but can continue to be added to and amended after its first release.

6. Concerns related to production

In terms of the production of the report, the main concern of researchers is that the medium should not assume more importance than the content. We are all aware that time could be spent on innovative production that could perhaps more usefully be spent on interpretation and analysis of the data. There are also concerns that, although the intention is to speed up the process by reducing production time, we may find that using a new production method would initially slow down those involved in producing the report because they were unfamiliar with the software or that new software would not always perform correctly. We were also concerned about producing both a paper product and a web-based product at the same time, which raised resource issues. Could we do both within agreed resources?

There are of course other concerns for the future about the use and interpretation of the data but this does not concern us in this paper.

7. The priorities

Having considered all the requirements, the priorities for the development of a web-based product were agreed as follows:

- produce within timetable
- produce paper and web product at same time
- initial features of web product to include: facility to search, download and print off copies of text, tables or charts; an electronic replica of the paper report; links between tables, text and charts; choice of view to include or exclude tables; links to a glossary; an index with links; basic

links to the user guide; the facility to view the product in proof-form and circulate it to clients for comment and amendments before it was generally available.

With these priorities agreed we moved on to look at the most appropriate technology to achieve them.

8. Multi Media publishing for ONS in general and Living in Britain in particular

Developing a multi media strategy for ONS, using the Living in Britain report as a pilot, required the recognition that if the information delivery specialists do their work properly then the specialist information providers (such as the social researchers) should not need to change profession and become HTML, ASP, JAVA or SQL programmers. They should be able to continue to use their expertise to collate, analyse and write expert commentary on the material while the mechanics of publishing this material should be 'transparent to them'.

This approach requires an integrated process for the creation and delivery to customers of statistical analyses and commentaries. This delivery is intended to be via a range of multi media processes. Those currently identified are:

- evolution of the current paper publications
- World Wide Web.

Other media such as DVD, WAP, E-Book or the 'next great invention' will all be accessible by adopting this multi media approach.

The basic premise is that the two component parts of ONS production need to be integrated. These are the commentary content - the expert analysis of the statistics - and the statistical databases - the actual numbers used as the basis for the expert commentary. Since Living in Britain contains both these elements the pilot project will be able to test both. This integration will be driven from different technology solutions.

Commentary and the process for presenting/displaying this information is controlled through a Content Management (CM) system, while statistical material is processed through central Information Management (IM) systems which support the full range of collection, collation, analysis and modelling. This data must then be available, either directly or through intermediate databases, to the CM system for use in either paper or web-based 'publications'.

9. Satisfying user requirements

Site infrastructure

The web site will be 'database driven' with pages of material displayed on screen being generated from the eXtensible Markup Language (XML) database using a variety of methods such as templates (Cascading Style Sheets), Active Server Pages (ASP), Dynamic HTML (D-HTML), Java Server Pages (JSP) and other web techniques, which will be transparent to the user.

The CM system will be able to supply the individual components of the web pages so that the basic design of the site will only contain a small number of database generated pages, which will vary their look and function according to the data being supplied from the CM system. This will ensure that configuration management of the site is kept to a simple level and that new views can be quickly implemented.

Customer profiles

It is possible for a module for allowing customer profiles to be stored to be included. This will ensure that when the web system is able to identify the user (either through cookies where permitted or user name/security as required) different levels of access, types of material and views of the site can be displayed for individual customers. The web site will continue to support 'anonymous' user access to a range of unrestricted material.

Customised views

Either by defining a specific user or by using a different access URL (e.g. www.statistics.gov.uk/student) or by selecting an option which sets the user view parameter, the system will allow differing views of, apparently, the same material.

One such example would be where the user parameter is set for a pre key stage 2 student (i.e. under 11 child), the site might display the material using a very graphical and 'wizard' style with inappropriate items from the database being filtered out (contraception, drug and alcohol abuse etc). If the user parameter were set for teacher then the view would probably be very different.

By having a custom view feature, the list of possible views is only constrained by the resources within ONS to build the templates for each view. The web system only sees these views as instructions in a database table and therefore entire new views can be generated very easily. This is becoming a very common approach with many web sites and is particularly appropriate for sites where multiple languages are supported since a each language is seen as a different view. A Welsh or Gaelic ONS site or even specific publications such as 'Living in Britain' in these languages are thus possible.

Secure connection

While the web site must continue to support 'anonymous' user access to a range of unrestricted material, the ability to secure elements of the site is important for issues such as advance release, pre-release proofing and specialist audiences. The combination of the customer profile and user passwords will allow for this functionality.

Document filtering

This feature is at the heart of making the service function effectively for the user. Every component held in the CM database will be 'tagged' with a variety of metadata indicating such matters as:

- title
- type of material
- themes covered
- keywords
- validity dates

- geographic/demographic/regional coverage

This metadata will be used by the web site to generate the relevant pages depending upon the individual requirements of the user at that point in their current session with the web site.

Users might start from a search page and select to only view methodological papers on a particular subject or statistics from a particular time frame covering a certain region. It would also be possible to select to view tables only or commentary only as required by different users of the GHS data.

Links to Statistical material

The CM system is not a statistical repository. It manages the links to 'resources' that are contributed from other sources/databases/folders/warehouses etc.

Within a particular article (or section of a report) stored in the CM the author might include the links to a graphic file (JPG of a pie chart perhaps) together with the links to the actual comma separated (CSV) spreadsheet file from which the pie data was compiled plus the link to a DTP version of the actual table (HTML, EPS, PDF etc). When assembling a web page of this article or section of the report, the CM system would supply the components and the web page might include the JPG of the pie chart on screen while also putting in an icon to link to the CSV file and another to link to the DTP table. Thus the user receives a web-optimised output, which allows them to drill down into other formats as required.

As the ONS Statistical Infrastructure evolves, the CM system will begin to link directly to pre-defined datasets in other databases so that the CM system simply holds the fact that table number 123456 from the statistical database is required at this point. The table would be generated by the statistical database, in the relevant format, at the time the request to view is made by the customer.

Interactive diagrams

Interactive diagram and on-screen mapping modules (server side objects that run on the web site) are important customer facing developments. These features are individual components, which will be referenced by the CM system and included into the web pages as required. Thus the CM system simply holds the fact that interactive diagram number 654321 is required to be inserted at this point in XYZ format. Development of these features will include their eventual connection to the statistical databases so that changes to figures in the database will be reflected by dynamic changes on the web site. This is an example of one of the many features that will be developed for Living in Britain after the initial release.

Printable/Downloadable versions

Individual components or articles will be held in the CM database. Many of these components link together to form complete products such as Living in Britain. Within the CM system there will be configuration tables that indicate that in product number 999 (Living in Britain) the component list starts with number 5674, then 8765, then 8890 etc to build up the entire product.

When viewed on the web site, the CM system will deliver the components in product order with indexing, 'skip to', links etc all built in and displayed as a web friendly page, in differing styles depending on the customer view currently set. This is an ideal format for browsing on screen but not for the customer printing to paper. Merely pressing the 'print' button on the customer's browser will customer printing to paper. Merely pressing the 'print' button on the customer's browser will not produce a 'paper friendly' version of the product. Therefore a separate PRINT option will be included which allows the customer to select how they want the product printed. Options would include defining the paper size, landscape or portrait mode, the inclusion or not of the graphic components, which chapters or pages to print etc. In this way the customer could choose to print out a 100 page product in a very similar version to a finished DTP product (see below), or they could print it as a summary only with just the commentary, or with all the tables but without the text.

Using a DOWNLOAD option the customer would also be able to use similar selection criteria as PRINT but be able to choose to have all or only some of the background data sets downloaded. This would be an ideal way for central libraries to download the product for loading to their own library network.

10. Satisfying requirements of ONS Staff

Individual creation of HTML pages by ONS staff will not be required. Contributors should be able to work through their current desktop computers using a suite of simple template style forms that load data to the CM database. It will be necessary to support levels of workflow hierarchy e.g. initial contributor - author - editor - approval.

With the growing move towards 'out working' or 'home working' plus the potential need for people in other departments to have access to elements of the CM, the system should be configured to allow access via the WWW with suitable user security. This would satisfy the requirement that the Living in Britain is accessible in its web format for commenting and amending by client departments.

Since every component held in the CM database will be 'tagged' with a variety of metadata values, ONS staff will need to indicate the types of metadata listed above.

11. Satisfying requirements for DTP

The CM system must support an output to DTP function so that publications can be 'compiled' within the CM from a list of existing components and then output in XML format to a DTP system. This will facilitate the concurrent production of a paper copy of the Living in Britain 2000/1 report. Prior to the publication, this system will be tested using the 1998 report, which has been converted to XML output. The entire publication will be processed through the chosen DTP package to test the functionality of the output routine. The statistical tables will be linked as individual resources. The CM system will incorporate the table reference at the indicated point in the DTP output (e.g. Table 5.6), any graphics (e.g. pie charts, pictures etc) and include the DTP version of the actual tables in the relevant place. Thus a paper product can be produced from the same system as the web-based product with little extra resource required.

12. Conclusion

Since the CM system is a database product, it can include all relevant types of material. For example:

- survey report

- news release
- exhibition summary
- methodological paper
- newsletter
- strategy document.

All such items are stored in the CM database and extracted/viewed in accordance with the filtering being applied by the product or customer using the database.

Thus, use of the Living in Britain 2000/1 report as a prototype for the development of a content management system satisfies all the basic requirements we have as researchers for a web-based system of dissemination for the results of the General Household Survey. Researchers must continue developing ideas for extending the range of possible features now that we have moved beyond the confines of paper products. However, we must also be aware that with this facility to continually change and adapt the outputs, we do not lose sight of our primary role as analysts and producers of data.

About the Authors

Alison Walker is a principal survey researcher in the Social Survey Division of the Office for National Statistics with responsibility for the General Household Survey. She has been involved in the paper publication of a number of survey reports but Living in Britain 2000/1 represents her first involvement with web-based dissemination.

Steven Connor is an independent consultant in the field of digital publishing currently working for the Office for National Statistics. He is a marketing man who became 'seduced' by the opportunites of multi media and has been involved in 'new media' since before the term was invented. He has worked with companies in the 'information' sector, trade and technical magazines as well as major book publishers both in the UK and internationally.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al

An Internet System for Clients: Continuous Survey Progress and Analysis Reporting on the Web

Rory MacNeill

Abstract

The Internet provides exciting opportunities for researchers. This paper considers the options for investment within a commercial market research environment. It describes the approach taken to focusing on the dissemination of survey information, both relating to the progress of a research project and to the results and analysis. It is highlights how the use of the Internet in these fields has generated benefits in other parts of the survey process. The paper concludes by considering some of the advantages that come from this approach.

1. Background

Using the Internet for research

In common with many other researchers, my colleagues I and have spent considerable time assessing the impact that the growth of the Internet can have on our working lives. We approach this from the position of being survey computing professionals interested in methodological advances but our plans have to be underpinned by the realities of commercial requirements and competition. The Internet presents exciting opportunities but also significant threats. Ones cosy position of expertise in the field of survey computing can be overtaken very quickly in new, fast developing areas.

We have questioned: What can the Internet help us do better? What is now feasible that previously could not be considered? What will we now do differently? What completely new services will now be required?

We invest in developing new services and I wanted to concentrate in areas that would benefit our customers most. Our process of deciding priorities went through the following stages:

- Clearly Web questionnaires were going to be useful in some situations and a necessary "tool" to have but we felt that they were less exciting technically and commercially than the "hype" would suggest. We provide this service but, unlike our competitors, we do not see it as a significant differentiating feature.
- Website evaluation was an interesting new area that our customers would want and that also became an obvious area to develop expertise.

- Other options considered included online panels, tracking web advertising and online focus groups but we felt that we should not depart dramatically from the type of work that we were already doing. We did not want to change our Brand position to use the Internet.
- Instead we decided to concentrate significant Internet development resources on what we do most of at present – telephone, postal and face-to-face interviewing.

Therefore, we concentrated our investment on developing on using the Internet for outward communication – the dissemination of information rather than its collection.

Using the Internet for telephone research

Once we recognised that traditional methods would continue, we set about considering how the Internet could be utilised within this framework. Very simply, a standard research project involves the following stages:

1. Design
2. Data collection
3. Analysis
4. Reporting
5. Assisting customers to use research results
6. A constant process of survey management and communication with stakeholders in the process.

A clear target for us was reporting research results via the Internet. Also, by providing better reporting tools to clients we could see an opportunity to improve usage of research results. This would not replace our consultants getting involved with business managers to help them make change but it would facilitate the process.

A secondary target was survey management and communication The Internet is ideal for such tasks. Whilst the benefits for reporting were obvious, the benefits from moving survey management and communication to the web were less clear but probably more extensive. In fact we began to see how easier and more extensive methods of communication could influence the design of the project to some extent.

So our development plan was set up on the basis that we would continue to collect a lot of information by telephone (or post or face-to-face interview) and continue to analyse it using database or statistical software. The Internet would be the bedrock for disseminating survey management information and analysis reporting with spin-off benefits in design and post research business impact.

2. Survey management

Recognising the opportunities

We wanted to make a difference to how research projects could be managed. To enable payback on investment we concentrated initially on large, complex projects. If the Internet can improve communication then we should use it in areas where communication is a major part of the project, that is clear. But how can we use the Internet to open up new opportunities within research projects?

We developed systems for clients with these two objectives in mind:

1 **Research where communication is a major issue.** This includes multinational multilanguage studies where a Web approach enables us to receive and provide information in a range of languages. Generally this leads to instant information exchange. At worst case communication can be translated by external bureaux and responded to within 24 hours. Another example where this approach could provide benefits is when there are many stakeholders in the research project. For instance: customer research where the customer relationship is managed by third party distributors or dealers; customer research where the customers are served by local outlets with some individual traits – retailers could fit into this category; or employee research where individual divisions of an organisation have different ways of managing their employees.

2 **Making different approaches feasible.** In the example above, the design of the research is sometimes restricted by what is considered feasible from a survey management and communication point of view. If survey management and communication is carried out on the Internet what can be done? Removing restrictions caused by project management limitations opens up new opportunities. For example, in an ideal world, a local manager would like to be able to supply details of the sample of customers (or staff) to be surveyed, the desired language for individuals, the preferred timetable, some "local" questions for specific individuals or groups of individuals and other information. In fact, this then lets him or her specify an individual piece of research within a general framework. Assuming that the manager is willing to pay for this degree of customisation then it is an ideal solution for the "parent" organisation also. From a respondent's perspective it is also beneficial. The respondent will be asked only questions that are relevant in a style and timeframe that is suitable. However, to date the effort in disseminating the survey requirements to individual managers and receiving their individual specifications and then implementing them has been too great a challenge.

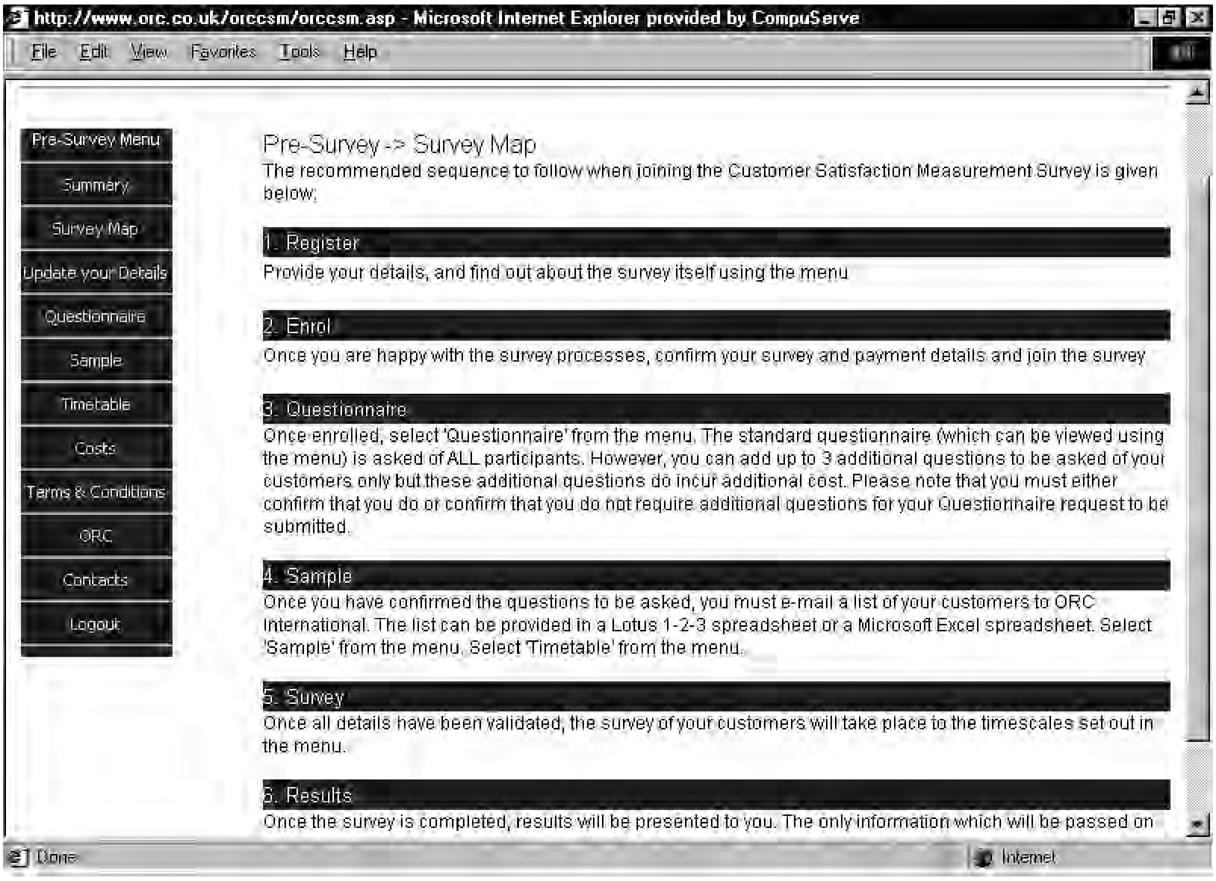
A client example – the IBM PartnerWorld programme

The appendix gives some detailed information on this research and our approach to it. The key element for this paper is the dissemination of information relating to the design and management of the research process. Our only contact with the representatives of the organisations participating in the research is through the website. Furthermore, often they will not be researchers; therefore it is crucially important that the information they are provided with enables them to specify a research project that we can then carry out.

The information that they receive from the website includes:

- Instructions on the process
- The core questionnaire
- Costs (for them to include customised questions)
- The timetable
- Contact information

- Terms and Conditions of their contract.



Pre-Survey -> Survey Map

The recommended sequence to follow when joining the Customer Satisfaction Measurement Survey is given below:

1. Register

Provide your details, and find out about the survey itself using the menu

2. Enrol

Once you are happy with the survey processes, confirm your survey and payment details and join the survey

3. Questionnaire

Once enrolled, select 'Questionnaire' from the menu. The standard questionnaire (which can be viewed using the menu) is asked of ALL participants. However, you can add up to 3 additional questions to be asked of your customers only but these additional questions do incur additional cost. Please note that you must either confirm that you do or confirm that you do not require additional questions for your Questionnaire request to be submitted.

4. Sample

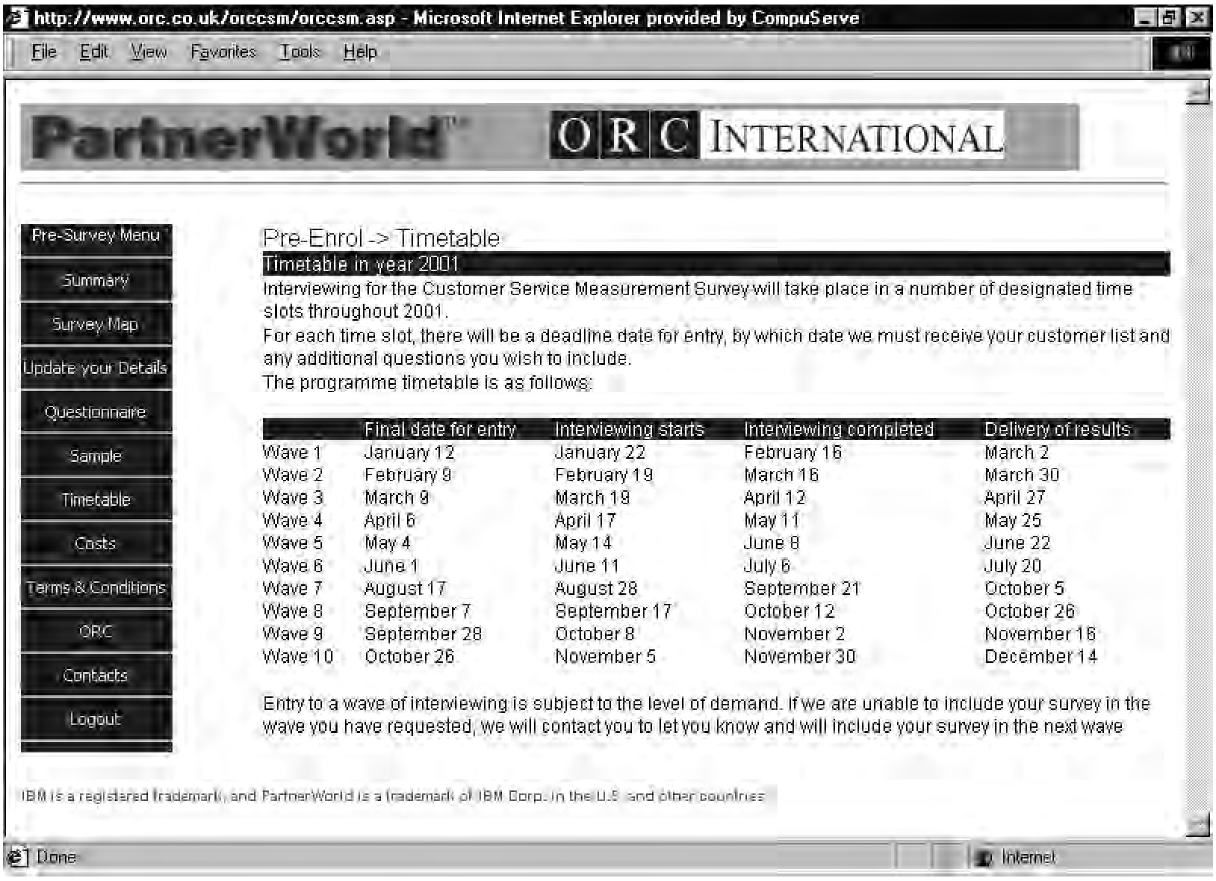
Once you have confirmed the questions to be asked, you must e-mail a list of your customers to ORC International. The list can be provided in a Lotus 1-2-3 spreadsheet or a Microsoft Excel spreadsheet. Select 'Sample' from the menu. Select 'Timetable' from the menu.

5. Survey

Once all details have been validated, the survey of your customers will take place to the timescales set out in the menu.

6. Results

Once the survey is completed, results will be presented to you. The only information which will be passed on



Pre-Enrol -> Timetable

Timetable in year 2001

Interviewing for the Customer Service Measurement Survey will take place in a number of designated time slots throughout 2001.

For each time slot, there will be a deadline date for entry, by which date we must receive your customer list and any additional questions you wish to include.

The programme timetable is as follows:

	Final date for entry	Interviewing starts	Interviewing completed	Delivery of results
Wave 1	January 12	January 22	February 16	March 2
Wave 2	February 9	February 19	March 16	March 30
Wave 3	March 9	March 19	April 12	April 27
Wave 4	April 6	April 17	May 11	May 25
Wave 5	May 4	May 14	June 8	June 22
Wave 6	June 1	June 11	July 6	July 20
Wave 7	August 17	August 28	September 21	October 5
Wave 8	September 7	September 17	October 12	October 26
Wave 9	September 28	October 8	November 2	November 16
Wave 10	October 26	November 5	November 30	December 14

Entry to a wave of interviewing is subject to the level of demand. If we are unable to include your survey in the wave you have requested, we will contact you to let you know and will include your survey in the next wave

IBM is a registered trademark, and PartnerWorld is a trademark of IBM Corp. in the U.S and other countries

Over 60 surveys were run in 2000, usually with no means of communication other than the website. So, clearly this type of approach can work. To extend the idea further the information received could be used to generate personal paper, CATI or Web interviews directly for individual respondents. These could be administered with practically no human intervention creating the opportunity to carry out personal research at low cost.

3. Analysis Reporting

Electronic reporting of research has been possible for several years. However, in our experience, paper reports have often been replaced with Adobe files of text, graphs and tables. This approach enables fast efficient distribution and printing but has very obvious limitations with regard to manipulating content. Within the market research industry there still seems to be a batch processing mentality to reporting: batches of tables and graphs are often provided for customers, written reports tend to cover every part of the questionnaire. Delivered reports are comprehensive and static. Their content is often specified before the data is available, and follow-up analysis is generally minimal due to fixed costs and timescales. If these working habits exist within customers of market research it is difficult to change them, however, how can the Internet be used to meet current requirements in a better way? How could other reporting styles that were previously infeasible be introduced?

Again, our client systems are designed to use survey data collected by any method. We report to clients via the project website using some or all of the following methods:

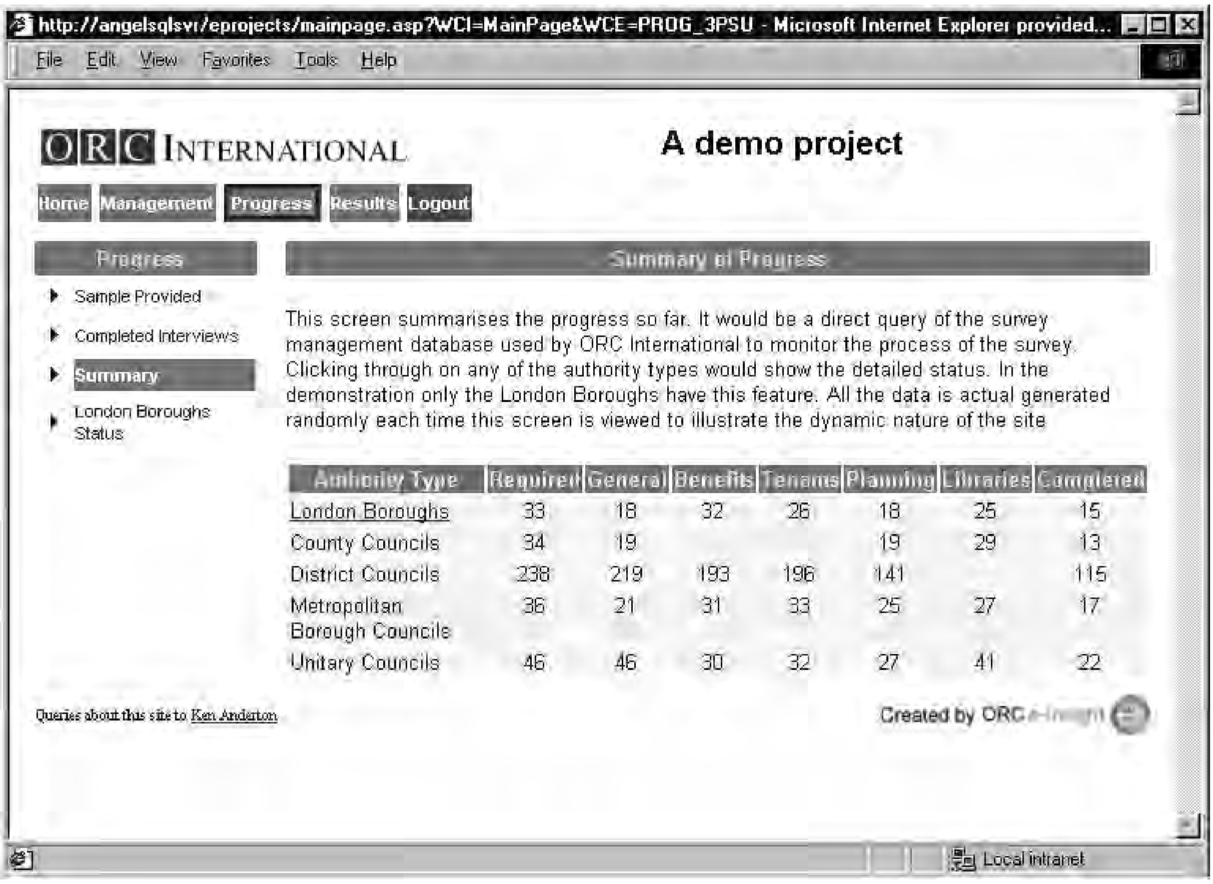
1. **Survey progress information.** Clients can view information on sample details, strike rates and survey process. This is updated daily from our CATI system for telephone research or our field and postal management systems. Also, information on issues and progress against project plan can be shown.

2. **Standard specified analyses.** Tables, charts and other figures will be designed and agreed with clients in advance of data collection. These can then be updated at appropriate periods during the research (daily, monthly or quarterly usually) and amended if certain features of the data are important.

3. **Flexible query systems.** Tables and charts that can be produced interactively across the Internet by customers in response to adhoc queries.

4. **Research library of written reports.** We have developed an application to hold all an organisation's written research reports on the Internet in an ordered way. This enables searches across all research reports and provides an efficient means of accessing information held in reports. The style of an Internet website is excellent for finding information from textual documents structured in an ordered way.

5. **Information on individual respondents.** In certain customer satisfaction studies it is extremely useful for client organisations to be able to identify what individual customers have said in response to survey questions. We update this section of websites very regularly with data collected by telephone, and this can be available on the Internet for clients within a few hours. This is very useful when high spending customers give negative comments. Obviously this type of research has to be designed carefully to comply with appropriate regulations regarding respondent confidentiality.


http://angelsqlsvr/eprojects/mainpage.asp?WCI=MainPage&WCE=PROG_3PSU - Microsoft Internet Explorer provided...
File Edit View Favorites Tools Help
ORC INTERNATIONAL
Home Management Progress Results Logout
Progress
Sample Provided
Completed Interviews
Summary
London Boroughs Status
Local intranet

A demo project

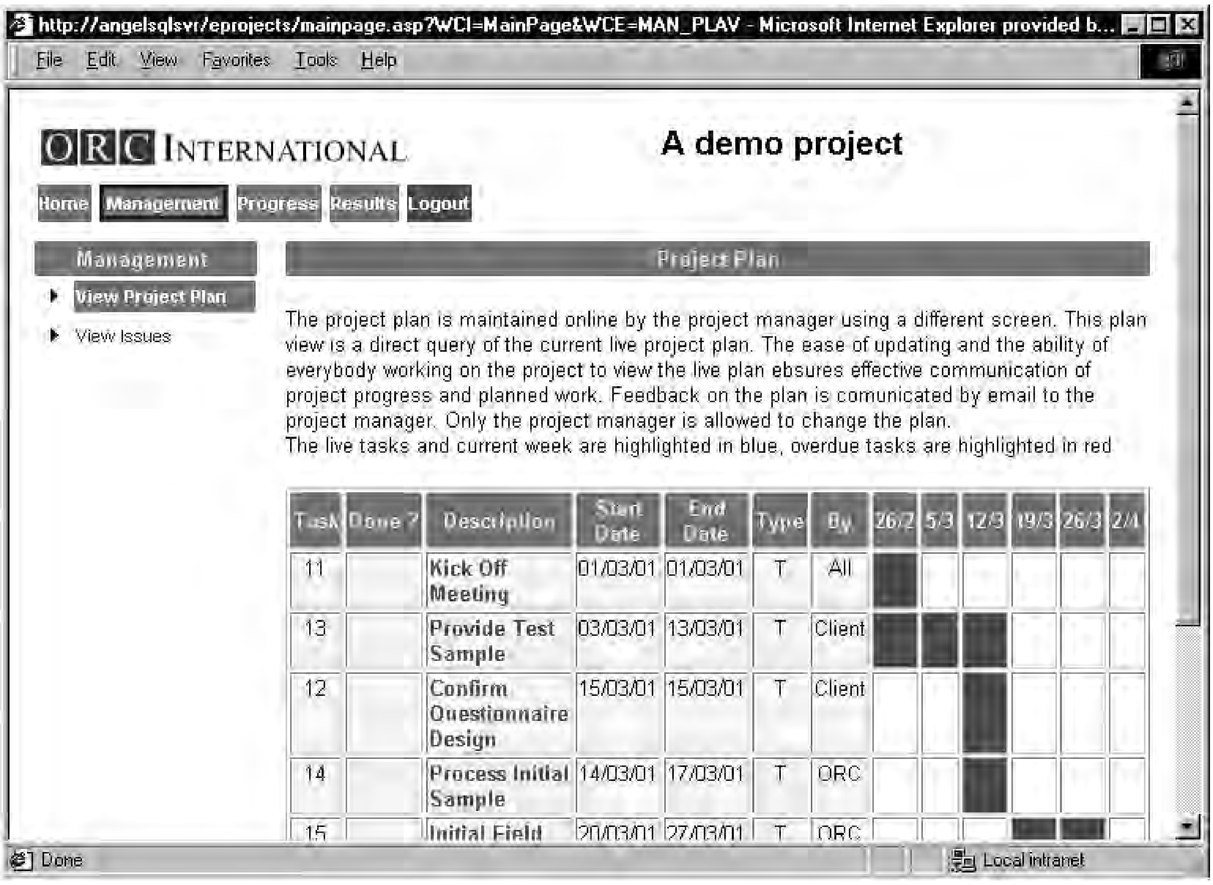
Summary of Progress

This screen summarises the progress so far. It would be a direct query of the survey management database used by ORC International to monitor the process of the survey. Clicking through on any of the authority types would show the detailed status. In the demonstration only the London Boroughs have this feature. All the data is actual generated randomly each time this screen is viewed to illustrate the dynamic nature of the site

Authority Type	Required	General	Benefits	Tenants	Planning	Libraries	Completed
London Boroughs	33	18	32	26	18	25	15
County Councils	34	19			19	29	13
District Councils	238	219	193	196	141		115
Metropolitan Borough Councils	36	21	31	33	25	27	17
Unitary Councils	46	46	30	32	27	41	22

Queries about this site to Ken Anderton

Created by ORC e-Insight

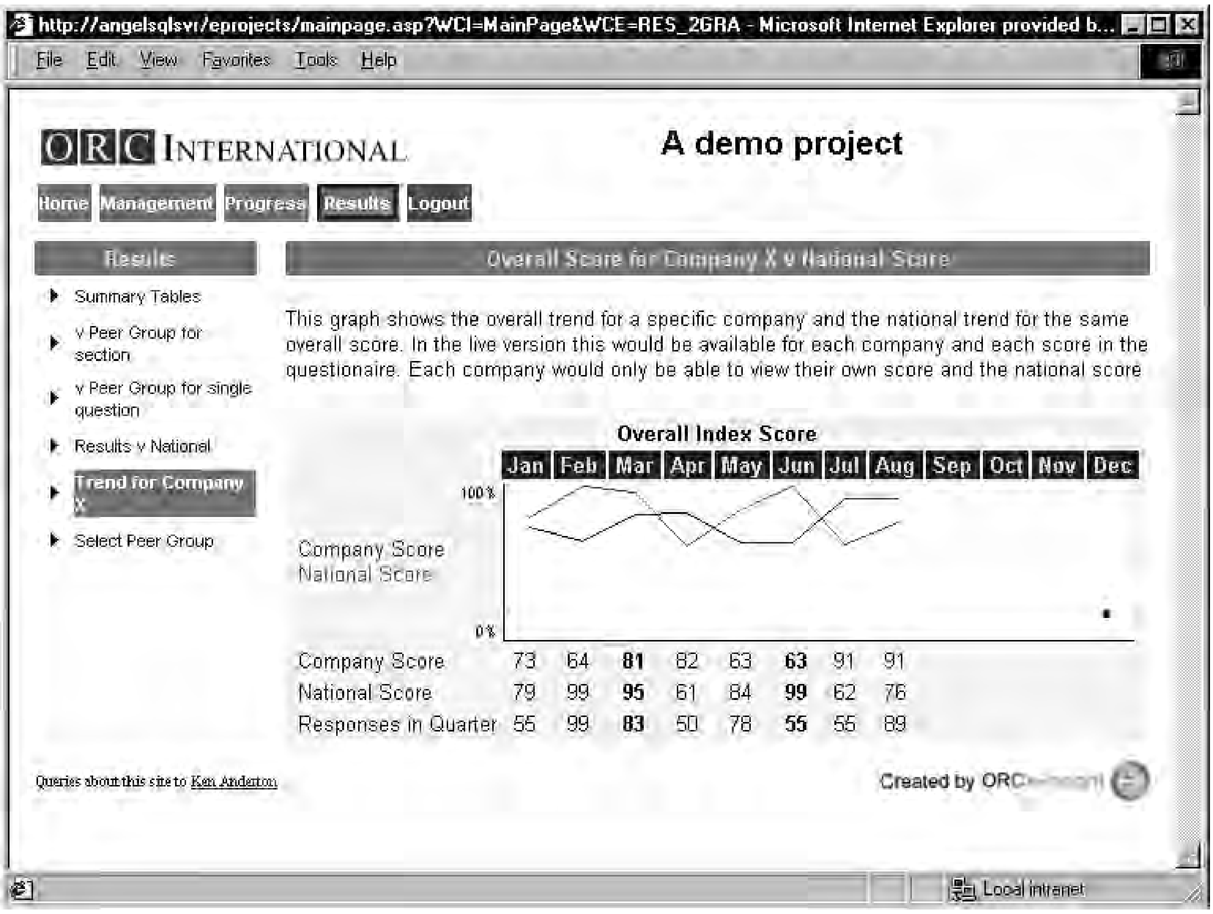

http://angelsqlsvr/eprojects/mainpage.asp?WCI=MainPage&WCE=MAN_PLAV - Microsoft Internet Explorer provided b...
File Edit View Favorites Tools Help
ORC INTERNATIONAL
Home Management Progress Results Logout
Management
View Project Plan
View Issues
Done
Local intranet

A demo project

Project Plan

The project plan is maintained online by the project manager using a different screen. This plan view is a direct query of the current live project plan. The ease of updating and the ability of everybody working on the project to view the live plan ebsures effective communication of project progress and planned work. Feedback on the plan is comunicated by email to the project manager. Only the project manager is allowed to change the plan.
The live tasks and current week are highlighted in blue, overdue tasks are highlighted in red

Task	Done ?	Description	Start Date	End Date	Type	By	26/2	5/3	12/3	19/3	26/3	2/4
11		Kick Off Meeting	01/03/01	01/03/01	T	All	■					
13		Provide Test Sample	03/03/01	13/03/01	T	Client	■	■	■			
12		Confirm Questionnaire Design	15/03/01	15/03/01	T	Client			■			
14		Process Initial Sample	14/03/01	17/03/01	T	ORC			■			
15		Initial Field	20/03/01	27/03/01	T	ORC				■	■	



A demo project

Overall Score for Company X v National Score

This graph shows the overall trend for a specific company and the national trend for the same overall score. In the live version this would be available for each company and each score in the questionaire. Each company would only be able to view their own score and the national score

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Company Score	73	64	**81**	82	63	**63**	91	91				
National Score	79	99	**95**	61	84	**99**	62	76				
Responses in Quarter	55	99	**83**	50	78	**55**	55	89				

Queries about this site to Ken Anderton

Created by ORC

Combining these means of disseminating information together with communication regarding survey management issues enables research consultancies to provide a fresh approach to partnering client organisations.

A client example – European customer research

A live client study illustrates some of the ideas described above. The organisation supplies high value consumer products and after sales service through third party distributors across Europe. They run a customer satisfaction programme measuring customer experiences of product purchase and after sales service. The information is collected through postal surveys (with a web alternative) in some countries, telephone surveys in others. The aim is to move the study to Internet surveying in the future, when access levels amongst the customer base improve.

The programme fulfils two roles: a measure of customer perceptions and behaviour and a measure of the performance of the distributors. Traditionally reporting had been through standard tracking reports produced monthly and quarterly at organisation, country and distributor level.

Utilising some of the techniques described in this paper has enabled the partnership between client and market researcher to change. The process becomes fresher, more instinctive and reactive rather than highly planned and mechanical. Instead of the routine arrival of statistics for a recent period, the website is updated daily and approaches to data presentation are flexible and easy to change.

In this example the client has available online:

1. **Sample information.** This shows progress of the survey, informing the client of the number of completed surveys, surveys in cleaning, loaded surveys and a summary of outcomes for each sample batch.
2. **Individual responses.** Information on each individual interviewed is shown. This links their survey results to the information supplied by the client as sample for the research. Each distributor can use this to look at detailed responses by their customers helping them manage their customer relationships. This information is available within one day of an interview.
3. **Alerts.** Where customers have made negative comments and requested a call from their representative this information is posted in a separate area of the website and updated daily.
4. **Survey results.** Data can be viewed in graphical (trend) format or as tables for chosen distributors, questionnaire sections or individual questions. This is available in a "drill-down" format to enable the organisation or distributors to identify where individual distributors differ in customer perceptions. This enables identification of areas of "best practice". Data is available from separate surveys for product sales and after sales service.

All of this is available in a choice of 11 languages.

4. Summary – what does this mean for the future?

Based on experiences of disseminating survey progress and analysis reporting using the Internet there are clear advantages for researchers:

Changes in approach

What was infeasible is now possible

The flexibility of the medium lets us do things that are quite obvious and sensible but were never considered previously because of limitations in time, budget and project management skills. We are now able to communicate survey requirements directly to hundreds or thousands of local managers, enabling them to specify local questionnaires, specific to their customers or employees.

Distribution issues become restriction issues

No longer do we have to consider who will receive distribution of research progress and results. It is now an issue of restricting access to those who can be trusted to use them. This is becoming a real issue for organisations keen to be "open" but shocked by how easy this is. If customer survey information includes strong criticism of individuals or small service departments should these be public to all employees?

Thus websites reporting information have to be designed with different levels of security and access control.

Think afresh on each project

Many of the processes undertaken in market research have their origin in the distant past but they still condition thinking in current methodologies. It is important to approach each research project with a fresh mind looking at what can be done rather than what has been done before.

Communication with clients is remote

There is a change in balance of communication. We now disseminate more information to more people but have far less personal contact. Most users of this information will be completely unknown to us. However, they do expect and are entitled to expect the same level of customer service as any other client.

Research companies are now e-tailers

Client organisations can now create commercial relationships with us over the Internet with no other contact. Our only positive communication with them is through an invoice and our own Customer Survey.

Business advantages

Customer Relationship Management

CRM is one of the important business topics within the marketing community at present and presents a major challenge to researchers. Research that enables flexible identification of small groups of customers helps in this field.

Survey fatigue

Designing surveys that are very local in nature, focussing on issues that are relevant to individuals reduces negativity to the survey process and research in general.

Adhoc reporting is available to all

In general, market research projects are reported on by an external agency. Often client organisations do not have the resources or skills to carry out an adhoc reporting once the business impacts of the research are being considered. If they do exist they tend to be central. Hence very little interrogative analysis is carried out after initial reports. Providing flexible online systems enables all decision makers access to the data.

Drawbacks

Ever increasing expectations

The challenge to satisfy will never cease. Although information is now available beyond what was ever previously considered possible, more is still expected.

Pressure on standards and codes of conduct

If market research can identify strengths of individual relationships in real time, client organisations are going to want this information. It is generally in the best interests of respondents for this to be available. However, current codes of conduct have to be negotiated. In the future the dividing line between research and individual consultation may be blurred as information on individual respondents is available in real-time.

Appendix IBM PartnerWorld programme

The PartnerWorld Programme

IBM serves most of its customers through independent third parties, known as Business Partners (BPs). Over the years, individual IBM brand and product organisations have created a number of marketing programmes to support these BPs and help them operate more effectively in the market-place. IBM is currently in the process of rolling out a new vehicle, PartnerWorld which will act as the single IBM marketing programme for end-user facing BPs.

The PartnerWorld programme offers numerous benefits to Partners, in the areas of marketing and sales support, technical support, education and training, financing arrangements, relationship management and, not least, financial incentives. These benefits are offered in a tiered manner to three different levels of BP: Member, Advanced and Premier. Partners qualify for the higher membership levels and their respective benefits by meeting a number of criteria that demonstrate commitment to IBM and investment of effort in building business for IBM.

One such criterion relates to customer satisfaction surveys. Advanced and Premier-level Partners are required to perform an annual customer satisfaction survey. Premier Partners are additionally required to obtain a set target score in the Net Satisfaction Index, which measures overall satisfaction with the BP among their end-user customers.

The Customer Satisfaction Measurement Survey (CSMS) Programme

Prior to 2000, BPs in IBM's Europe, Middle East and Africa (EMEA) region who wished to carry out a customer satisfaction survey were responsible for conducting or commissioning one on their own. For 2000, IBM wished to establish a research programme, which a single research supplier would be responsible for administering on its Partners' behalf.

This was seen as offering a number of benefits, both to the Business Partners and to IBM itself.

Firstly, Premier Partners, who are required to achieve a target score in the survey, have a right to expect that other Premier Partners will have been judged by the same criteria as themselves. With Business Partners conducting their own research, the result was that inconsistent, and sometimes dubious, methodologies were used. Using a single reputable research supplier and a standardised questionnaire helps to ensure consistency and legitimacy. End-user customers also have a right to expect that IBM judges all its Business Partners by the same standards, and that accreditation can be trusted as an impartial and consistent recognition of a BP's competence.

A further benefit is that the use of a standardised set of core questions enables each Partner to benchmark its performance against that of others and IBM, to have a common global or regional standard against which to monitor its BPs' performance.

Finally, there are obvious economies of scale in combining all the BPs' individual surveys into a single programme. By spreading set-up and management costs across a large number of surveys, IBM was able to obtain a price for each survey, and for the total programme, that was far lower than could have been achieved had each Partner commissioned its own research. This helped not only the BPs, who make the initial outlay for the survey, but also IBM, who agree to reimburse the cost of the survey through one of the elements of the PartnerWorld programme.

The Research Challenge

The CSMS programme was not especially demanding in respect of its data collection and analysis requirements. However, with regard to survey management, the design of the programme raised a number of exceptional issues that required special preparation and logistics.

- Firstly, the decision to participate in the programme rests solely with the Business Partners themselves. Partners are free to choose whether or not to seek the higher levels of PartnerWorld membership that require them to conduct customer satisfaction research and this requires that an 'on-demand' service should be made available to them. Similarly, the timing for embarking on the research process lies entirely at the discretion of the BP. Although IBM may set deadlines for achieving PartnerWorld membership criteria, these deadlines do not necessarily dictate when BPs decide to conduct their surveys. Thus the programme needed to provide Business Partners with continuous access both to **information** about the survey and to the processes required for **enrolment** to the survey. It was anticipated that the Business Partner would initiate enrolment by completing some type of application form, which would then be forwarded to the research agency. Whilst mailings to Business Partners might be useful in communicating initial information and subsequent reminders about the survey, it seemed unreasonable to expect BPs to retain any hard-copy documentation until such time as they needed it in order to enrol.
- Applications were expected from Business Partners in more than 20 countries in Europe and beyond. Any communication to BPs would need to be available in a range of languages.
- Additionally, the programme was required to provide the facility to handle ad hoc queries – again likely to be delivered in a wide range of languages – within an acceptably quick turnround time.
- Of critical importance is the fact the Business Partners are in competition with each other – and with IBM itself, which can also sell direct to end-users. Thus, Business Partners maintain their own customer lists and are highly protective of them; there is no central source of IBM's end-user customers. The customer lists would need to be supplied by the Business Partners themselves and these would have to be kept confidential. BPs needed to have complete confidence that appropriate security procedures would be in place to prevent unauthorised access to the customer lists. The reliance upon BPs to provide their own customer lists also raised the possibility that BPs might seek to manipulate the sample, 'cleansing' it of customers with whom they had experienced problems, in order to try and improve their Net Satisfaction Index score.
- The programme had to offer the potential for Business Partners to add extra questions to the standard questionnaire, in order to address issues of particular interest to an individual BP. This was felt to be particularly relevant to those BPs who were already running surveys of their own and wished to continue with certain questions from their current surveys. However, this process would need to be carefully managed in order to avoid a massively complex set-up process.
- A wide range of survey deliverables was required. BPs were to be supplied with graphical summaries of their results as well as individual comments from those customers who had agreed to have them passed on. In addition, on completion of the annual programme, BPs would receive another graphical report, comparing their individual results against those from the programme as a whole. At the same time, IBM needed to receive ongoing reports listing the BPs who were participating, the number of customers on their sample lists (to provide IBM with an

opportunity to assess whether BPs were being selective in the customers listed), and their NSI scores.

- Last but not least, a large number of Business Partners from different currency zones each had to be invoiced individually and a simple and secure payment system had to be established.

Finding a Web-based Solution

On seeking to meet the challenges described above, it was evident that a web-based solution was highly appropriate. Indeed there seemed to us to be a happy coincidence between the needs of this research programme and the impulses that stimulated the origination and development of the Internet and the World Wide Web.

The application of computer-aided communication was originally developed within the intellectual community. The reasoning behind it was that, for any significant problem, there would be only a few people who could contribute effectively to its solution. These people were likely to be widely dispersed, highly independent-minded and often not the best team players. At the same time, some form of intellectual partnership, in which ideas could come into contact with each other, was clearly essential.

The vision behind the creation of the web was that of a common information space, in which people could communicate rapidly and share information without being brought physically together in one place or, indeed, without compromising their independence. In its infancy, the Internet was used only as a medium for broadcasting read-only material but more recently, of course, it has been used in an ever-increasingly interactive way.

The compelling case for using the web for the CSMS Programme was the need to maintain ongoing two-way communication with the several hundred IBM Business Partners who represented the potential purchasers of the service. Given the nature of the market, it was taken as read that all the Partners would have Internet access. A purpose-built website, dedicated to the CSMS, was therefore constructed to provide an easy entry point for Business Partners into the CSMS process.

For the initial launch of the website, ORC International prepared an introductory letter, which was sent out to Business Partners, explaining the enrolment and survey processes and giving the web address. From this point onwards, reference to the web address became a familiar part of the regular communication between IBM’s local offices and their Partners.

The Process

On first visiting the site, Partners are required to register, giving a user name and a password name. At this stage, Partners are required to give some basic details including their organisation’s name, a contact name and an e-mail address. The list of registrations is passed back to IBM to provide a check on the bona fides of the Partner, providing some safeguard against the unlikely event of a non-authorised organisation locating and entering the site.

At the registration stage, Partners are not committing themselves to participate in the research programme. However, by giving us their e-mail address, Partners enable us to start a dialogue with them. For example, shortly before the start of each interviewing ‘wave’, we can send out reminders to Partners who have not followed up their initial registration to see if they are now ready and willing to send in a customer list and proceed with the survey.

The site's primary function, at the registration stage, is to provide information about the survey. Partners are able to find out details including the list of standard questions, definitions of the type of customer to be interviewed, the timetable for the survey, costs of participation, payment terms, information about ORC International, as well as a link to a dedicated e-mail address that has been established for the survey.

Naturally, because of the wide geographical spread of the survey, the information on the site has to be made available in a number of languages. Our decision was to set up the site in five languages – English, French, German, Italian and Spanish – which was expected to make the site easily usable by all Partners. On entering the site, the first screen brings up a set of flags that enables Partners to select the language of their choice.

After registration, Partners are able to enrol for the survey on-line. At enrolment, some further details are requested:

- Partners are asked if they require any extra **questions** additional to those contained in the standard questionnaire. If so, they are directed to a form on which closed questions can be entered, with space for the question text and a set of responses. Using such a form enables us to control the format in which questions are submitted, although all questions need to be 'vetted' by ORC International to ensure that they are sound in research terms.
- Partners are required to send in their **customer lists**. The website contains an example of how this should be laid out, to ensure that we receive all the detail that we require. Partners are able to download this template and send us the customer list via the secure e-mail link contained on the site.
- We also ask for details of the **currency** in which Partners require to be invoiced. To make entry to the programme easier, Partners are given the option of paying in Sterling, Euros or US Dollars.

The fieldwork itself is conducted by telephone. We attempt to contact all of the Business Partners' customers (usually fewer than 100) within the designated fieldwork period and, so far, have achieved interviews with around 45% of all the customers submitted.

Behind the scenes, the website is linked to a survey management database that stores all the subscriber information centrally, serving as a record of all those who are interested in the CSMS programme and allowing us to track the progress of each subscriber. This enables us to provide regular progress reports to IBM as well as triggering the despatch of reminder e-mails to Partners, to encourage registered Partners to send in customer lists - or to chase up payments!

A key factor in the project's success is the use of a dedicated e-mail address to which the whole of our CSMS project team has access. This has two major benefits. Firstly, by using e-mail as the main medium of communication with Business Partners, we are able to deal easily with a multi-lingual subscriber base. Queries e-mailed to us in languages other than English can easily be translated 'off-line' and replied to (in the native language, if necessary) within 24 hours. By contrast, running a multilingual telephone helpdesk would pose a much more severe logistical challenge. In addition, the use of a 'project' address, rather than an individual address, enables us to provide continuous cover, no matter which members of the team are present.

References

Traditional vs. web. B2B panel research in the Internet age, ESOMAR 2000
Robert McKane, Manager, Bell Atlantic, USA
James Heisler, Executive Vice President, Opinion Research Corporation International, USA

The benefits of Web-based survey management for a global dealer network, BIG Conference 2001
Mike Joseph, Research Director, ORC International

About the Author

Rory MacNeill is a director of ORC International, previously having been a Director of SIA Ltd prior to acquisition. His current role includes responsibility for developing and marketing ORC International's Internet research services through the company's ORC e-Insight brand.

He has worked in the field of survey computing for 18 years. He has worked on many client projects in the public and private sector over this time, applying survey computing techniques to research issues.

Previous ASC papers include:

Evaluating CAPI for the Family Resources Survey, 1993 (jointly presented with Stuart Mitchenall, DSS)

Analysing Open Ended Questions, 1996 (jointly presented with Ken Anderton, ORC International)

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

Using Agent Technology to Disseminate Statistics via the Web

Karen Brannen and Joanne Lamb

Abstract

MISSION (Multi-Agent Integration of Shared Statistical Information Over the [inter]Net) is a multinational project funded by the European Commission. It aims to provide a modular system of software that will enable providers of official statistics to publish their data in a unified, and unifying, framework, and to allow consumers of statistics to access these data in an informed manner with minimum effort.

The project is co-ordinated by the University of Edinburgh, with partners in the Office of National Statistics (UK), the Central Statistics office of Ireland, Statistics Finland, the University of Athens, the University of Ulster and DESAN Market Research of Amsterdam.

Statistical Offices have in recent years been focussing their attention on the World Wide Web as a method for information dissemination. This focus is exemplified by the development of StatBase in the UK and StatLine in the Netherlands. In addition discussions at the UN/ECE workshop on dissemination to the media has increasingly looked at the implication of web based dissemination for contact with journalists and other consumers.

To date, however, this activity has been approached from a somewhat centralist viewpoint. Either the disseminating agency is concerned only with its own material, or it has collected information from a variety of sources and made it available from a single site. Neither approach is viable in the longer term. Users will want to be able to obtain information from several sites, and will look for an efficient way of gathering the information from these sources. The overhead in collecting data from several sources and locating it at a central site will become prohibitive.

The MISSION project takes a different approach. It aims to permit distributed heterogeneous data sources to be available to a user through access to a Library which holds the details of available datasets. Libraries consult each other, so once a user has accessed one Library, he or she has access to all data sources registered in a MISSION Network.

Since MISSION is a web based system, it can be integrated with existing web sites that run the normal functions of advertisement, news releases, contacts and customer services.

This paper will further describe the aims of the project and achievements to date in the development of the MISSION system and how it can contribute to developments in the dissemination of statistical information.

Keywords

Official statistics, heterogeneous data, dissemination, agent technology

1. Introduction

MISSION (Multi-Agent Integration of Shared Statistical Information Over the [inter]Net) is a multinational project funded by the European Commission. It aims to provide a modular system of software that will enable providers of official statistics to publish their data in a unified, and unifying, framework, and to allow consumers of statistics to access these data in an informed manner with minimum effort.

The project is co-ordinated by the University of Edinburgh, with partners in the Office of National Statistics (UK), the Central Statistics office of Ireland, Statistics Finland, the University of Athens, the University of Ulster and DESAN Market Research of Amsterdam.

Statistical Offices have in recent years been focussing their attention on the World Wide Web as a method for information dissemination. This focus is exemplified by the development of StatBase in the UK and StatLine in the Netherlands. In addition discussions at the UN/ECE workshop on dissemination to the media has increasingly looked at the implication of web based dissemination for contact with journalists and other consumers.

To date, however, this activity has been approached from a somewhat centralist viewpoint. Either the disseminating agency is concerned only with its own material, or it has collected information from a variety of sources and made it available from a single site. Neither approach is viable in the longer term. Users will want to be able to obtain information from several sites, and will look for an efficient way of gathering the information from these sources. The overhead in collecting data from several sources and locating it at a central site will become prohibitive.

The MISSION project takes a different approach. It aims to permit distributed heterogeneous data sources to be available to a user through access to a Library that holds the details of available datasets. Libraries consult each other, so once a user has accessed one Library, he or she has access to all data sources registered in a MISSION Network. The aims of the project and methods used will be further described later in this paper.

Project Objectives

The World Wide Web has already had a profound impact on the way NSIs publish data. By the end of the Fifth Framework Programme, in 2003, Europe will truly be working in a global market and providers of official statistics will have to work in this context. This has a number of implications:

Users, governmental organisations, businesses, education and private citizens, will have access to a wide range of statistical datasets from a diverse, and largely unregulated, group of providers.

As access to statistics becomes easier, users will wish to compare the statistical information produced from these datasets and to perform comparative analyses on datasets that have not been harmonised by an authoritative body.

The demand for statistics, the ease of access, and the rate of change in economic and social conditions will require a dynamic sharing of expertise by domain experts and statisticians in order to address the social, economic and policy issues which will arise.

The vision of MISSION is a number of independent organisations publishing their data within a framework that makes comparisons and harmonisation possible.

The objective of this project is to provide a modular system of software that will enable providers of official statistics to publish their data in a unified, and unifying, framework, and to allow consumers of statistics to access these data in an informed manner with minimum effort.

We shall utilise the advances in statistical techniques for data harmonisation, agent technology, the availability of standards for exchanging metadata and the power of Internet information retrieval tools, to build a modular software suite. In this vision, the suite will:

- allow suppliers of statistics to subscribe to an integrated network of datastores via an interface to their existing data while retaining control over all aspects of access to their data: their level of involvement; the data they supply; the users who can access it; and the level of resources to commit;
- allow users to make declarative requests, with a minimum of understanding of statistics, or the domain area, and still retrieve meaningful results from our internal routines or through an interface with external statistical packages;
- give the user a range of options for "automatic harmonisation" of statistical data, with clear indication on the interpretation of the results;
- provide audit trails of data manipulation and analysis, so that methods can be retained, re-used and published;
- maintain libraries of metadata which can be made available to other users;
- provide a flexible architecture which allows third parties to act as Independent Metadata Providers, thus encouraging the free exchange of knowledge;
- allow users to build up individual profiles, accessing data and methods most relevant to their needs;
- offer a number of independent, interoperable systems which can run on different hardware platforms and access heterogeneous data storage systems.

2. Methods

The software suite will be made up of three basic logical, or conceptual, building blocks. These units can be deployed in a number of ways, thereby ensuring a flexible system. The units are the Client, the Library, and the Data server. Each unit will be described in a little more detail before showing how they are used together to form the system.

The Client

This is an application downloaded from the Web, which connects a user to a host MISSION site (a Library). The client is a Java application that provides the GUI interface to the MISSION site. This interface is described below.

The Library

This is a repository holding statistical metadata. In the project proposal we identified three different types of statistical metadata, all of which are held in the library. *Access* metadata is the most basic type which is the physical and logical information required to access statistical data. *Methodological* metadata is the information required to process data in order to satisfy requests for statistical analysis. *Contextual* metadata supplies background information and explanatory notes for the user. This broad classification is used in the context of the functionality of the MISSION system. Access metadata is imported with the data. Methodological metadata is generated by the creation of results from request. Contextual metadata must be read by the browser to help the user determine the usefulness of the data.

Libraries communicate with each other, thus giving the user access to all Libraries.

Each library holds a statistical processing engine.

The Data server

This is the unit that gives access to the data. Each data server holds the data itself, management tools for registering and maintaining the system and a gateway module. The gateways hold the minimum amount of metadata necessary for the safe use of the data, including registration details to allow the data provider to control access to the data and information about the physical structure of the datastore. Each data server connects to one or more libraries.

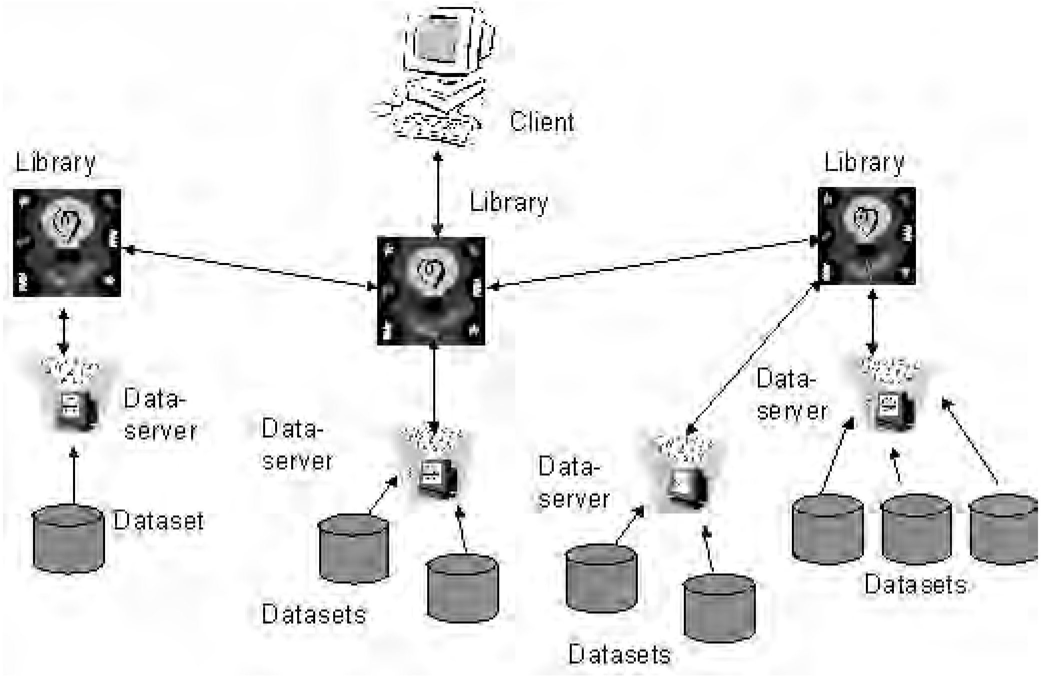
Agents

The units described above form the static components of the whole system. The use of *agents* to develop a peer-to-peer environment is one of the innovative methods used within the MISSION project.

The static units operate using generic computations and access or store various data repositories. However, the user orientation of the system is carried out using agents. These agents perform intermediate processing and navigate the Internet to access the appropriate unit. Once the unit is located and accessed, agents are responsible to invoke the appropriate computations on the engines or retrieve the appropriate data and metadata according to the request.

Figure 1 shows a general picture of a typical MISSION network.

Figure 1: a typical MISSION network



3. Concepts

In this section of the paper, we describe in a little more detail some of the main concepts used in the MISSION system. The Table-driven user interface examples draw heavily on work done by Karel Pagrach and Nico Caarls of DESAN Market Research of Amsterdam.

The Table-driven user interface

DESAN Market Research of Amsterdam, one of the MISSION consortium partners, is developing the idea of the table-driven user interface. Most users think of data in terms of output, more specifically in table form as in Figure 2. The particular example shows courses, with percentages shown at an institutional level (inst) as well as the national level (nl). The column entitled *Aantal resp.* shows the number of valid cases. The table shows two variables, each with two categories, but the first variable only shows the data for one category (Deelname vervolgonderwijs). Experience shows us that when users receive such a table, their first question is generally how can the table be changed in order to tailor it to their own needs.

Figure 3 shows a specific example of a table of results that is sent to individual institutions. Examples of the small changes that can be requested from such a report are shown here. They include merging cells, adding new variables or inserting or splitting a variable.

Figure 2: Example of an output table

TABEL 3.7 VOLTIJD

DEELNAME AAN EN NIVEAU VERVOLGONDERWIJS

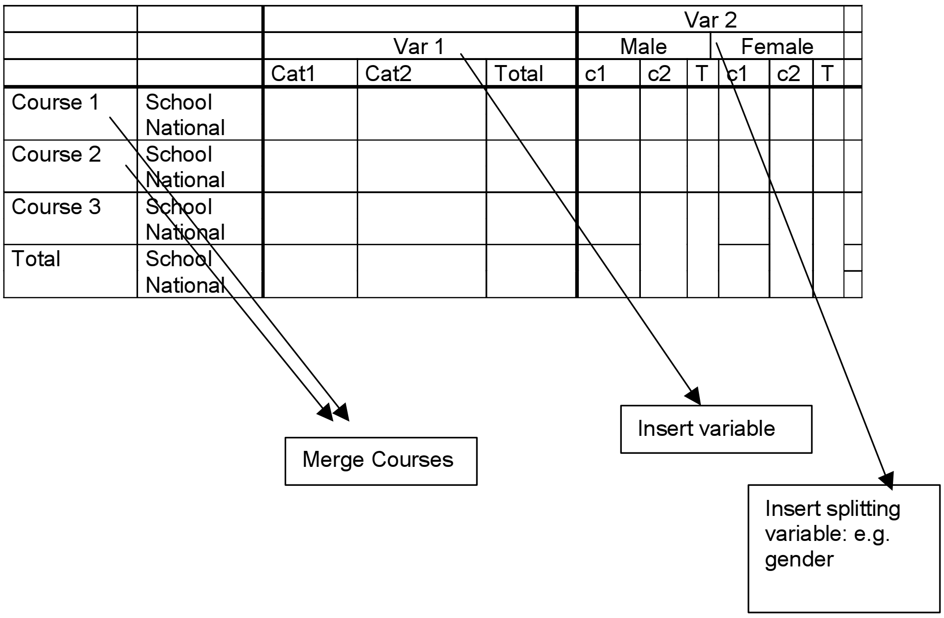
		Deelname vervolg-onderwijs	*Aantal resp.*	HBO	WO	*Aantal resp.*
HBO Accountancy	inst	39%	*46*	27%	73%	*15*
	nl	38%	*208*	24%	76%	*64*
HBO Management, economie en recht	inst	28%	*67*	17%	83%	*18*
	nl	27%	*441*	11%	89%	*118*
MET DIPLOMA	inst	*33%*	*113*	*21%*	*79%*	33

VOORBEELD HBO

Ideally, the user should be able to make these changes interactively to an on-line version of the table, as illustrated in Figure 3. Taking this one step further, the MISSION project aims to incorporate an interactive table definition into its user interface, thus allowing the user to not only change the existing layout of a table, but to define the required layout.

Defining a layout and style on the one hand and a query on the other hand are usually seen as two different tasks, which should be done in two different ways. However, the two tasks are indeed related. It is difficult to define a layout without also defining the related query. It is therefore a further aim of the MISSION system to use the same interactive method of defining table layout, to allow the user to define the required query.

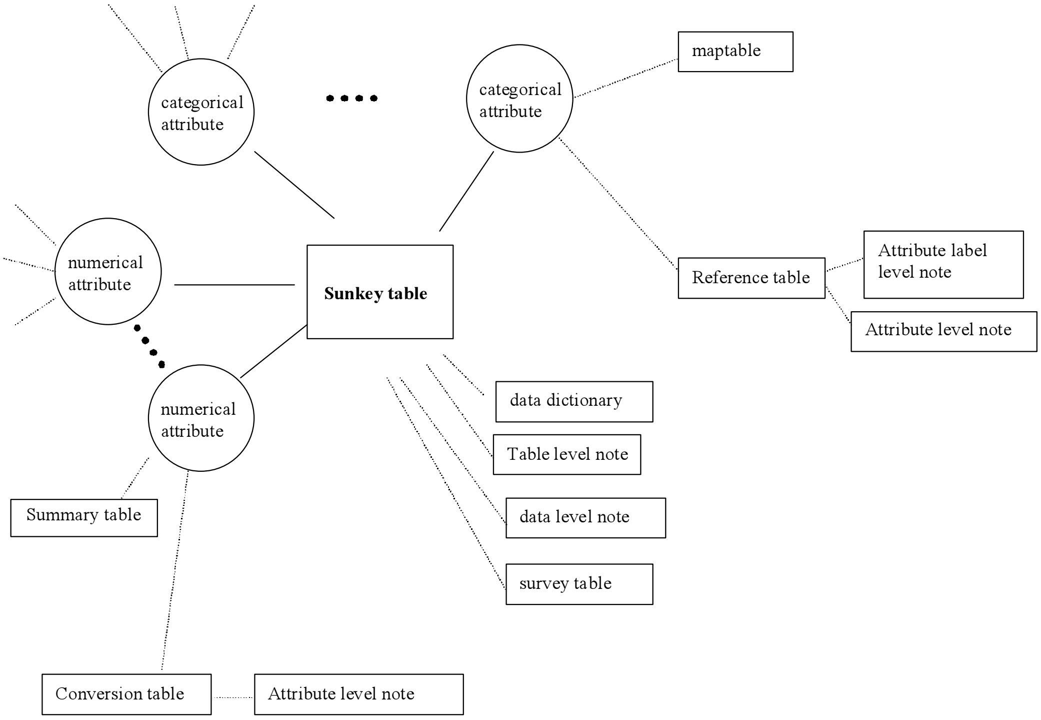
Figure 3



MIMAMED model

The MIMAD (Micro-Macro Data) model, previously developed at the University of Ulster (Sadreddini *et al.,* 1990, 1991, 1992a, 1992b) and as discussed in Scotney & McClean (1997), was extended in the precursor project to MISSION (ADDSIA1997) to a MIMAMED (Micro-Macro-Meta Data) model to take account of the active metadata that is involved in statistical query execution.

Figure 4



A number of National Statistical Institutes (NSI's) envisage the adoption of data warehouses in the future for both regular statistical production and knowledge discovery. Statistics Netherlands has already developed a standard for electronic statistical publication called StatLine. This table-based standard is currently being revised to a cube-based model (van Bracht & Sluis, 2000).

Taking this into account, the MISSION project aims to adopt a MIMAMED model that can interact with data warehouses and can exploit current developments. It will also allow for metadata accommodation. In order to do this, a hybrid model will be developed in which a data cube may be transformed into a MAMEOB (macro-meta data object).

Libraries

Libraries have been briefly described earlier in this paper. The main concept here is to allow a distributed network of metadata stores. Every library in the network provides a common interface to the client. Users are free to connect to any library they wish and, through queries sent to that library, access the entire metadata structure. The library consists of a number of sections, which are described below. These sections are logically separate and interacting but not necessarily physically separate.

The access metadata facilities include the location and structure of the data available in the system. It also holds the permissions regarding which users can access the data as well as what they are permitted to do with that data.

The methodological metadata facilities include the storage of the methods used to obtain a particular result from the raw data present in the system. This metadata can be examined in detail to show the specific operations performed by the system in order to audit and verify the result obtained.

The context metadata facilities include the storage of all background metadata necessary to put the stored data in context. This information often appears as footnotes to tables. An important facility provided by the context metadata section of the library is the ability to search on a wide range of terms. This can help the user to frame a suitable query.

The metadata mapping facilities include the storage of the mappings between the raw, source data and the system or user wide ontologies that are created or developed during the lifetime of the system. Further mappings can be derived to extend and complete the mapping network.

The library agent's facilities include the ability to dissect user queries and search the network of libraries for a suitable method to answer the query. Library agents also manipulate the mapping metadata to derive new mappings as well as ensuring that all the metadata is consistent and arranged for efficient access.

Within the Library is a statistical processing engine, which stores no information of its own. Based on the query it receives, it obtains the necessary data from various data servers, performs the request, and returns the result to the Library that made the request. It may also make a request to third party statistical packages.

The facilities of this processing engine include the ability to take data and metadata from different heterogeneous sites and merge them according to different criteria selected by the user. The module also offers an interface to external statistical packages, to provide specification and software to receive and interpret a user-specified query from a library unit. It will contain: query agents that utilise appropriate library units and appropriate data server units; statistical macrodata operators and associated metadata operators in order to answer queries, and mediation agents to enable user-specified merger of heterogeneous macrodata and metadata. It will also generate query plans that provide new action metadata for future query optimisation.

Transformations

When analysing data there are a number of steps that are taken. Having found the datasets and the variables in them, we go through a cyclic process as described below.

First we analyse the basic data to identify what is there. We then reshape the data to fit our hypothesis of how the world should be - this may be a research hypothesis, or a standard indicator. Having done this, we might tabulate this reshaped form of the data. These last two processes are reiterated until we are satisfied that the results we are producing reflect the question we are trying to answer.

We can identify a number of different methods used to transform data. Usually a number of input or dummy variables can be defined, with a single result. Different methods have different requirements. When a transformation is applied, the variables must have the same characteristics as the dummy variables of the transformation. Below we list some of the different methods that can be used in a transformation.

Method	Comment
Computation	uses an arithmetic expression to construct the result
Truncation	Removes one or more digits from the end of a code - usually for a classification
Rule-based	a series of instructions of the form IF xxx THEN result = y
Banding	a specification that a range of values should be collapsed to a single code
Transcoding table	a table of initial and final results
Complex	a combination of methods

Each method has its own set of relevant metadata, which allows us to specify an algorithm for the transformation. Suppose we want to recode a variable of five divisions into two main divisions. It is the decision of the user that values 1 and 2 (say) fall into one category and the rest into another. He therefore defines a new transformation and a new codelist. At the same time as defining the transformation and codelist, the user identified the reason for the transformation. By giving the user the opportunity to define why (or when) a transformation should be applied as well as how it is done, we capture important metadata for the documentation of indicators.

Publishing 'methods'

A key idea of the MISSION system is the sharing of methods – the transformations described above. It is often the case that data cannot be made freely available. In many cases data is only available under specially licensed circumstances, which means that any modifications made to the raw data under (say) a research investigation are not available to other investigators who want to follow up previous work. While recognising this restriction, the MISSION project aims to encourage the sharing of expertise by allowing users to publish the methodology that they have developed when working with data through the MISSION system. This means that other users can apply the transformations identified by subject matter experts in their own analysis, and so benefit from the shared expertise. The requirement of fully documenting the transformations or methods that are published is therefore self-evident.

4. Progress to date

The MISSION project is at the beginning of its second year. The first year has been concerned with developing the conceptual model further, collecting user requirements (Pagrach *et al*, 2001) and testing the connectability of the system. In addition we have developed the definitions of the agent functionality, particularly in the area of finding and matching suitable data. The project plans to develop two prototypes. The first prototype will demonstrate the specification and retrieval of a simple query. The second will incorporate the full functionality, which allows the user to browse methodologies and tables that have been created using the MISSION system. These resources will be held as XML files generated by the system.

Discussions with users identified a need for an interactive table specification as described in section 3. This user interface is designed for simple users with no knowledge of analytical techniques. By using drag and drop techniques, they can specify the descriptive final table as they wish it to appear. This visual request is translated into a declarative request using the following query language (Froeschl, 1997).

Operator	**Operand**	**Example of operand**
COMPUTE	Table	Table, graph or model
OF	N	n, mean or s.d.
ON	Frame	Survey, eg LFS
FOR	Target concept	Numerical attribute eg salary
WITH	Predicate	SEX=Female
BROKEN-DOWN-BY	Cross-product of categorical attributes	GENDER by JOB
OVER	Geo-referenced categorical attribute(s)	EU-Countries
IN	Temporal-referenced categorical attribute(s)	YEAR =1990 thru 1999

This declarative statement is processed in the library using a number of query agents. These query agents together carry out the task of analysing and processing the request. They can be described as follows:

A Matching agent receives a rough query (attributes for groups or frames), and also receives the required classification schemes. It matches the query against metadata by matching schema. It interacts with the *negotiation agent* to determine a match that is compatible. It interacts with the *covering agent*, passing information on the appropriate datasets and ontologies to be used in the query.

A Negotiation agent matches classifications/ontologies. It receives classification schemes from a matching agent. It determines if the classification schemes can be matched to a global ontology (by interaction with classification servers), and returns information to the matching agent.

A Covering agent receives the query, datasets and classification as determined by the matching agent. It determines which combination of datasets can optimally cover the query. (For example, in the simplest case, determine if all attributes for a group are in a single dataset. Use nulls if some of the attributes are missing.) It interacts with the *costing agent* to determine the cheapest suitable datasets. It passes on covering information to the plan generation agent. An example of how this dataflow works is shown in the following diagram.

Once the query has been analysed, it is passed to the query planning agent, which decides how to split the query into sub queries, and then sends request to the information agents sitting at the different information stores.

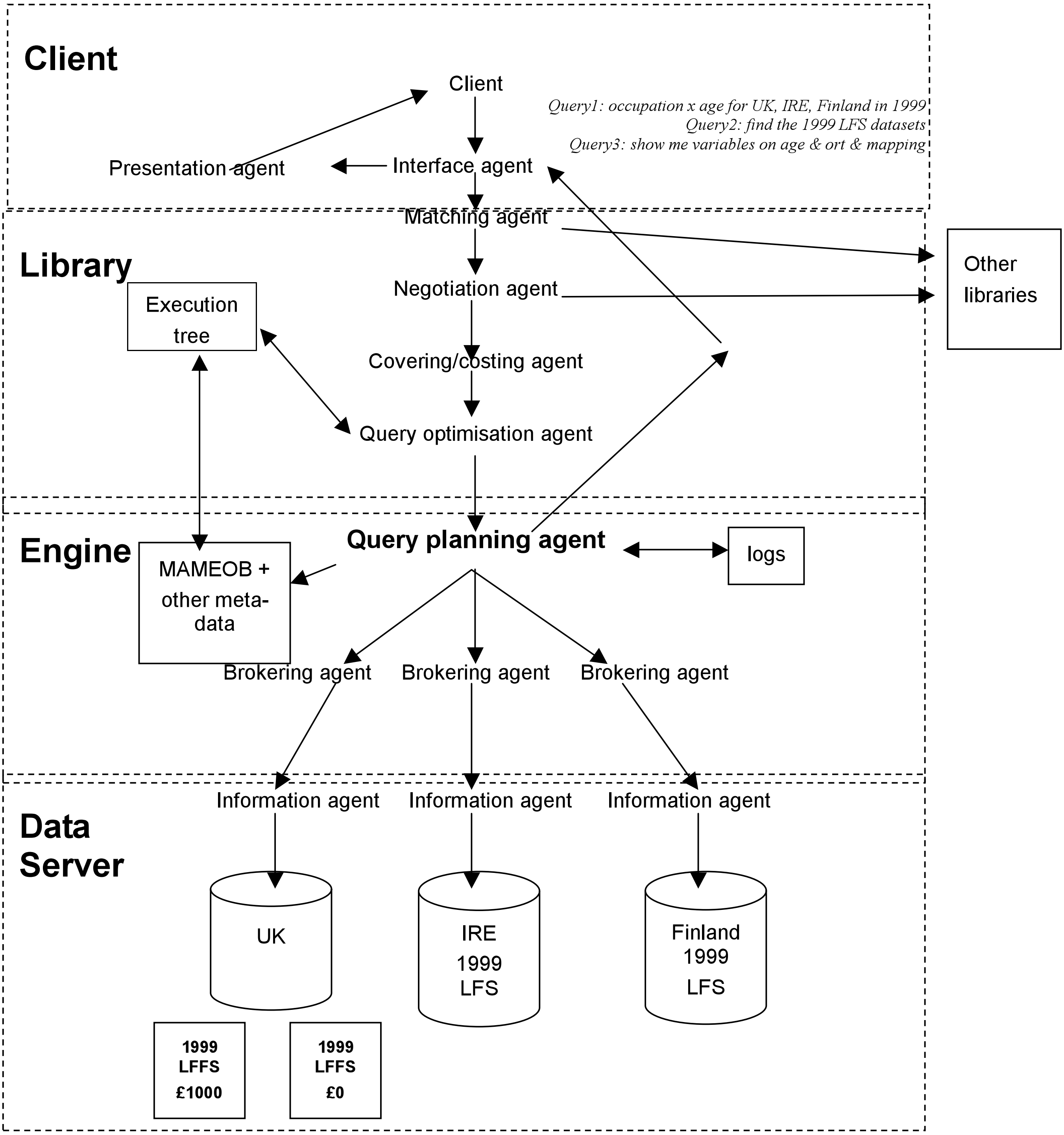
A simple version of this architecture was successfully demonstrated at the end of February 2001. The first prototype, due for September 2001 is targeted to. Its functionality will be

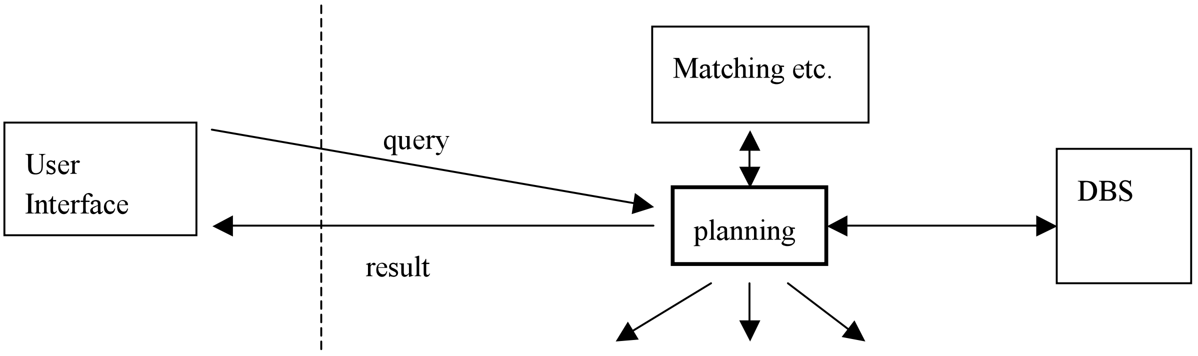
- Browse metadata – ie view dimensions of an aggregated dataset
- Produce 'simple' tables – ie data from a single source
- Produce 'composite' tables – ie data from different sources, but with identical dimensions, ie with a mechanism for provider to provide mapping (1 to 1) at classification level.

The data used will be the World Health Organisation Mortality Data Base (MDB) which is a compilation of official data on causes of death and population reported by Member States to WHO annually.

5. Conclusions

The MISSION project aims to satisfy a number of different but important objectives. From the technology point of view we seek to use agent technology to reduce the burden on users in identifying suitable datasets and data. We also introduce a user-friendly graphical interface to query definition and table modification. For more advanced users we offer the opportunity to formalise and publish their methodology thus sharing expertise. A key feature in this approach is the open and distributed nature of the system, where Libraries can exist independent of the dataset owner, and where owners are free to make their data available through a shared network of resources, while still retaining control of the access rights to their data.


Client
Client
Query1: occupation x age for UK, IRE, Finland in 1999
Query2: find the 1999 LFS datasets
Query3: show me variables on age & ort & mapping
Presentation agent
Interface agent
Matching agent
Library
Execution tree
Negotiation agent
Other libraries
Covering/costing agent
Query optimisation agent
Engine
Query planning agent
logs
MAMEOB + other meta-data
Brokering agent
Brokering agent
Brokering agent
Data Server
Information agent
Information agent
Information agent
UK
IRE 1999 LFS
Finland 1999 LFS
1999 LFFS £1000
1999 LFFS £0


Matching etc.
User Interface
query
planning
DBS
result

References

Annex 1: Users' Guide to Standardized Transcripts on Electronic Storage Media, in 1994

Access to Distributed Databases for Statistical Information and Analysis, EU, ESPRIT, Jan 1997 - Dec 1999

Froeschl, K.A. (1997) Metadata management in statistical information processing, New York: Springer Wien.

Pagrach. K., Rutjes, H. and Sluis, L. (2000) "Synthesised description of user needs" Deliverable 4 of the Mission project.

Sadreddini M.H., Bell, D.A. and McClean, S.I. (1990) "Architectural Considerations for Providing Statistical Analysis of Distributed Data", *Information and Software Technology*, Vol. 32, pp.459-469.

Sadreddini M.H., Bell, D.A. and McClean, S.I. (1991) "A Model for Integration of Raw Data and Aggregate Views in Heterogeneous Statistical Databases", *Database Technology,* Vol. 4, No. 2, pp.115-127.

Sadreddini M.H., Bell, D.A. and McClean, S.I. (1992a) "A Framework for Query Optimization in Distributed Statistical Databases", *Information and Software Technology*, Vol. 34, No. 6, pp.363-377.

Sadreddini M.H., Bell, D.A. and McClean, S.I. (1992b) "Providing Statistical Functionality in a Distributed Environment", in A.Westlake, R.Banks, C.Payne & T.Orchard (eds) *Survey and Statistical Computing*, North Holland, pp.467-476.

Scotney, B.W. and McClean, S.I. (1997) "Using Database Technology to Facilitate Statistical Analysis of Distributed Data", New Techniques and Technologies for Statistics II, IOS Press, Amsterdam, pp.203-213.

Scotney, B. (2000) "A Micro-Macro-Meta Data Model for MISSION" WP4.2-no5-UU.doc, Internal Working Paper, Ulster: University of Ulster.

Scotney, B. (2000) Library Agent Hierarchy WP4.2-no6-UU.doc, Internal Working Paper, Ulster: University of Ulster.

Van Bracht, E., de Jonge, E. & Kaper, E. (2000) "CRISTAL Data Objects. An Object Model for Cubic, Rar, or Intermediate Statistical Data". Netherlands: Statistics Netherlands.

Van Bracht, E. & Sluis, W. (2000) "Towards an International Standard for Multi Dimensional Tables", Netherlands: Statistics Netherlands.

World Health Organization World Health Statistics Annual. Geneva: WHO.

About the Authors

Karen Brannen

Karen Brannen has a BSc (Hons) in Computer Science, Aberdeen, and is currently a Research Fellow at CES. She is responsible to the Assistant Director for the provision of Information Systems and Information Technology Support within the Centre, as well as participating in research projects. She was responsible for data management and maintenance of the SYPS series of surveys conducted by CES, and has managed the datasets for the CATEWE (EU TSER SOE2-CT97-2019) project, and other comparative analysis projects. She has participated in the IDARESA (ESPRIT 20478) project, and has written a number of papers on the practical issues of harmonising data. Her research interests include accessibility and use of computing technology for research purposes; methods of improving data quality, accessibility and storage; use of data communications in the field of academic research; automation of survey processing; harmonisation of data from different sources.

She is currently working on several research projects: MISSION (IST-1999-10655); IQML (IST-1999-10338) and METANET (IST-1999-29093). The MISSION consortium consists of:

- University of Edinburgh, Scotland, UK
- Office for National Statistics, UK
- Central Statistics Office, Ireland
- Tilastokeskus (Statistics Finland), Finland
- University of Athens, Greece
- University of Ulster, Northern Ireland, UK
- Desan Marktonderzoek BV, The Netherlands

Dr Joanne Lamb, Assistant Director of the CES

Dr Lamb is currently project co-ordinator of the MISSION project (IST-1999-10655); the IQML project (IST-1999-10338) and the METANET project (IST-1999-29093). She also leads the CES contribution to AMRADS (IST-2000-26125). She also participated in the DOSES project EISI. She has a PhD in Computer Science, and she was Survey Manager of the Scottish Young People’s Survey for a number of years, before becoming Assistant Director of the Centre. Dr Lamb has developed a questionnaire design and management package, a Metadata Management System linking questionnaire design, data capture and data documentation and a general Survey Administration System.

Integration through Software and Metadata

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al

The Effect of Standards on Software Component Architecture

Chris Nelson

Abstract

There is a tendency for the specification of software interfaces to migrate from tightly coupled APIs, such as COM and CORBA, to loosely coupled interfaces based on "messages". In parallel with this, and acting as an important facilitator for this, is the explosive use of Internet technology. The international standards making process has facilitated greatly the ability of small software companies to develop robust and functional software. By following, or even contributing to, this process a software developer can construct state of the art software in a fraction of the time (and therefore a fraction of the cost) it used to take. By developing open standard interfaces, these software components can be integrated to build bigger and more functional systems.

Keywords

Standards, XML, OMG, UML, ebXML, IQML

1. A Week is a Long Time

A British Prime Minister once remarked, when asked about a change of opinion compared with a statement made the previous week, that "A week is long time in politics". If a software developer of the same epoch had made the same sentiment about the technology when asked why the software design had been changed, he would have been ridiculed. However, times have changed and while a week is still a long time in politics, but it is now also a long time in software technology evolution. The Internet (and the standards it has spawned) has caused an explosive growth in the choice of technology available to implement a particular system. And it is changing by the week. So, how can the software architect of today design systems that will not only be state of the art, but also give maximum protection against the ever onward march of technological development? Do standards have a role to play? Can we learn from the technology adopted or used by the standards organisations in the software industry?

The answer is yes to both questions. It is interesting to see how others approach software architecture design and we can learn a lot and save time by adopting a similar approach.

In order to demonstrate this, I will discuss how the IQML project, with which I am involved, has taken advantage of the standards making process, both in its own design and its feedback to the process.

1. The IQML Project

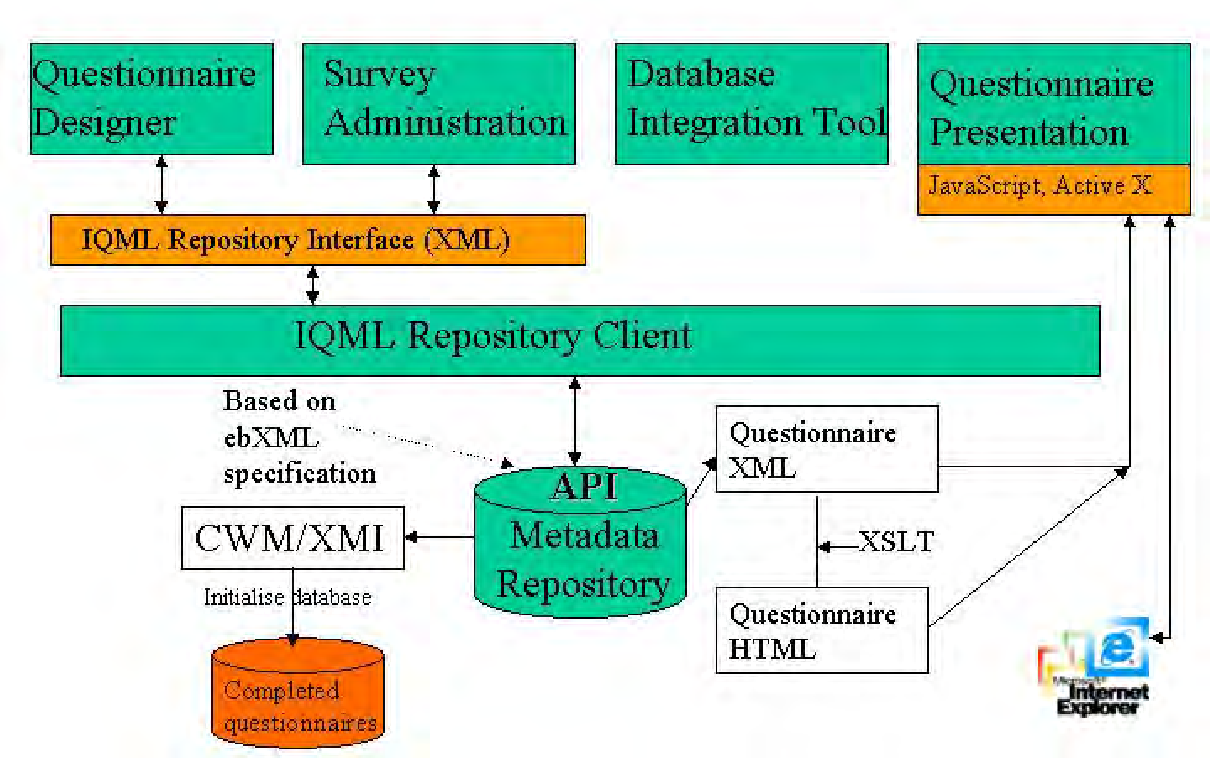
IQML stands for Intelligent Questionnaire Markup Language. It is part funded under the EC 5th Framework programme. There are seven partners – 1 university, 3 users (including another university), 2 software houses, and 1 survey company. Four of these organisations are developing software.

Two of the main goals of the IQML project are:

- To support the EU in its quest for more accurate and timely information whilst reducing the burden on enterprises that supply this information, by using current and emerging information technologies to implement a solution for intelligent questionnaires.
- To ensure the metadata interchange and database access standards being elaborated at the international level by software vendors (eg OIM from the metadata coalition, CWM from the Object Management Group) takes into account the needs of the intelligent questionnaire, by participating in the standards process and by developing products, and re-engineering existing products, that use these standards in live data collection scenarios.

The software schematic of the IQML product portfolio is shown below.

Schematic of the IQML Products



The intent of the project team is to use open standards wherever possible, and to take advantage of the standards process to ensure the architecture is consistent with the way the large software companies are thinking. The standards processes that have been followed are those that involve the software houses, and so the project hopes to adopt standards that will lead to software interoperability.

In addition, by being an active member of the standards making process, the project is contributing to the standards made in ebXML and the OMG. These acronyms will be explained in the course of this paper.

2. Model the system

The object model gives stability to the specification

The first thing we did on the project was to model the system. The reason for this is that whilst the technology used to implement a system will change, perhaps dramatically and perhaps over a short period of time, the logical model of the system should be quite stable over a much longer period of time: the fundamental dynamics of the business does not change very much. So it makes sense to model the system and to then represent that model in one or more syntactic implementations, adding new syntactic representations as the need arises.

In any system design it takes much longer to agree the fundamentals of the system in terms of functionality and logical structure, than it does to specify the technical implementation. It is worth spending a large percentage of effort on this logical design. As we will see from the CWM, the actual technical specification is done mechanically from the UML model. The UML model for CWM took two years to develop: and this was done by experts in the domain and experts in UML meeting frequently. The XML specification took a matter of minutes.

In the field of aggregated statistical exchange between statistical organisations and in the Central Banking community, it is the model that is the key. The model took a number of years of to develop (but this was at a much slower rate of development than the CWM). The technical specification was hand crafted in EDIFACT but this took only a matter of days. An XML specification of the same model can be hand crafted in less than one day.

In the IQML project, the model of the Question Bank, Questionnaire, and Survey Administration has taken many months to develop, but the technical design of the Repository to store and access the underlying objects has taken just a matter of a few weeks, and was actually documented in just 5 days. Detailed technical design of the underlying processes is ongoing but the model has allowed us to specify all of the interfaces and the data store.

One thing to understand about an object model is that you don't need to implement the model as objects and you don't need to implement the data store as an object oriented database, or even a relational database – simple flat files, such as comma separated or XML, can be used to implement the data storage mechanism.

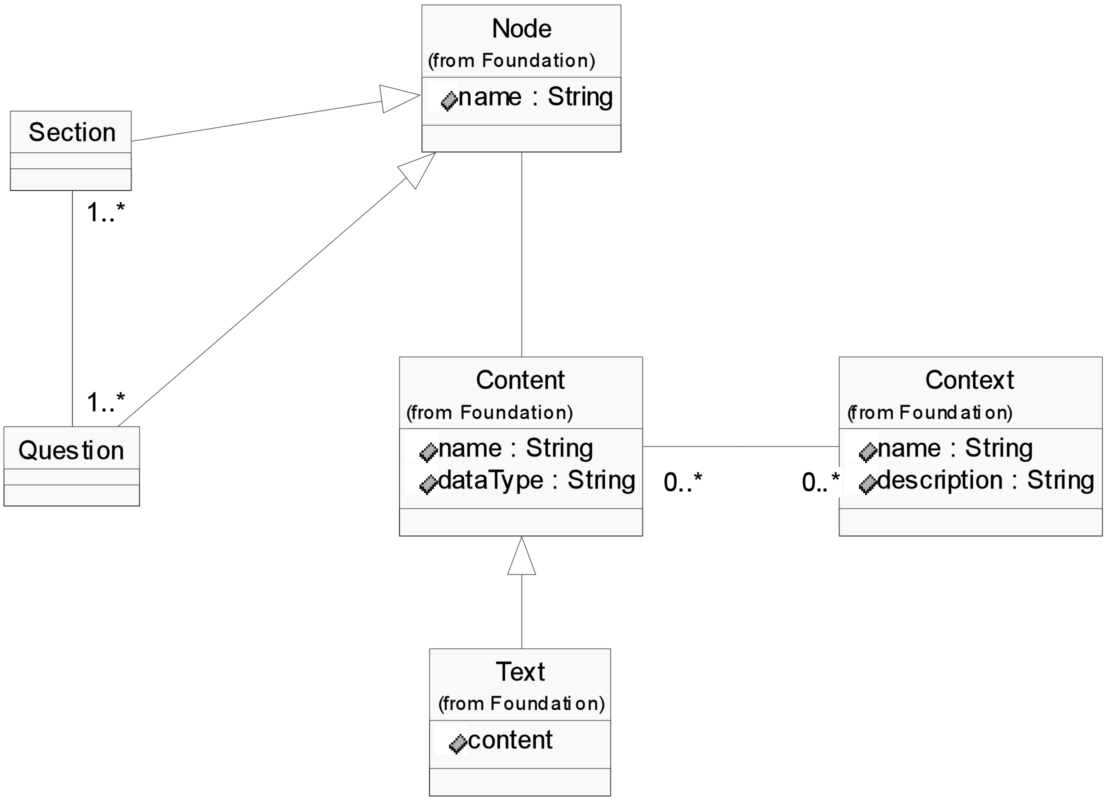
The accepted standard for object modelling is the Unified Modelling Language (UML). Why unified? Well, UML was born out of an amalgamation of three leading modelling methods (Booch, OMT, and Objectory). This amalgamation was carried out within the OMG technology process and the UML is an OMG standard and is maintained by the OMG. UML has been adopted by ebXML and UN/EDIFACT as the modelling language to be used to specify the logical system.

UML is a graphical language created for specifying, visualising, constructing, and documenting the artefacts of a software system. Some UML tools will generate large parts of the application code if system has been specified in sufficient detail in the UML.

However, many projects use the UML as a means of specifying the data required and the interfaces, and are therefore concerned principally with the logical view of the system – often called the data model.

The more abstract you can make the model the more flexible it will be when it is implemented. The degree of abstraction is important: too much abstraction and the software developer has a lot of work to do, too little abstraction and the system will be inflexible and difficult to maintain. Models of questionnaires need to be at a certain level of abstraction because the model must deal with all types of questionnaire and all types of question.

A small extract from the IQML Question Bank model is shown below.



This diagram shows that any Node can have Text (Text is a "type of" Content). The Text can be classified by Context. Context can be anything – age, gender, language, nationality, text type etc. So, it is possible to classify any Text by multiple classifications (e.g. question text in French for age group 10-20). Most objects in the Question Bank are sub classes (i.e. "type of") Node. Therefore, both Section and Question can have Text. A code label is also supported this way – the Code is a "type of" Node and the Text can be given the Context of "Label". Multi-lingual labels are supported by also giving the

Text a language Context. There are other “type of” Content that are not shown on this extract (e.g. Image).

When modelling, it is useful to see how others have solved similar problems. You can even buy books of UML patterns. For the IQML project, we were fortunate to have been involved in development of the Object Management Group (OMG) Common Warehouse Metamodel (CWM), where techniques similar to the one used above were used. If such a model can be implemented, then you can see how flexible the system can be.

3. Implementing the model

The repository

In order to illustrate the advantage taken by the IQML project of another standard, it is interesting to look at the implementation of the repository.

On the IQML project we decided to have a single repository which stores all the metadata that is to be shared by the various applications, or that is required by a downstream process (e.g. post collection process). An application can interface to the repository to generate the questionnaire in a format compatible with the collection instrument. The major role of the applications that use the repository are to act a Graphical User Interface (GUI) between the user and the repository. So, for instance, the Questionnaire Designer is really a GUI sitting on top of the repository, enabling users to design a questionnaire from a bank of re-usable questions, and to add/modify questions in the bank

The IQML model can be implemented in many ways, but the approach taken follows closely the work done on the ebXML process.

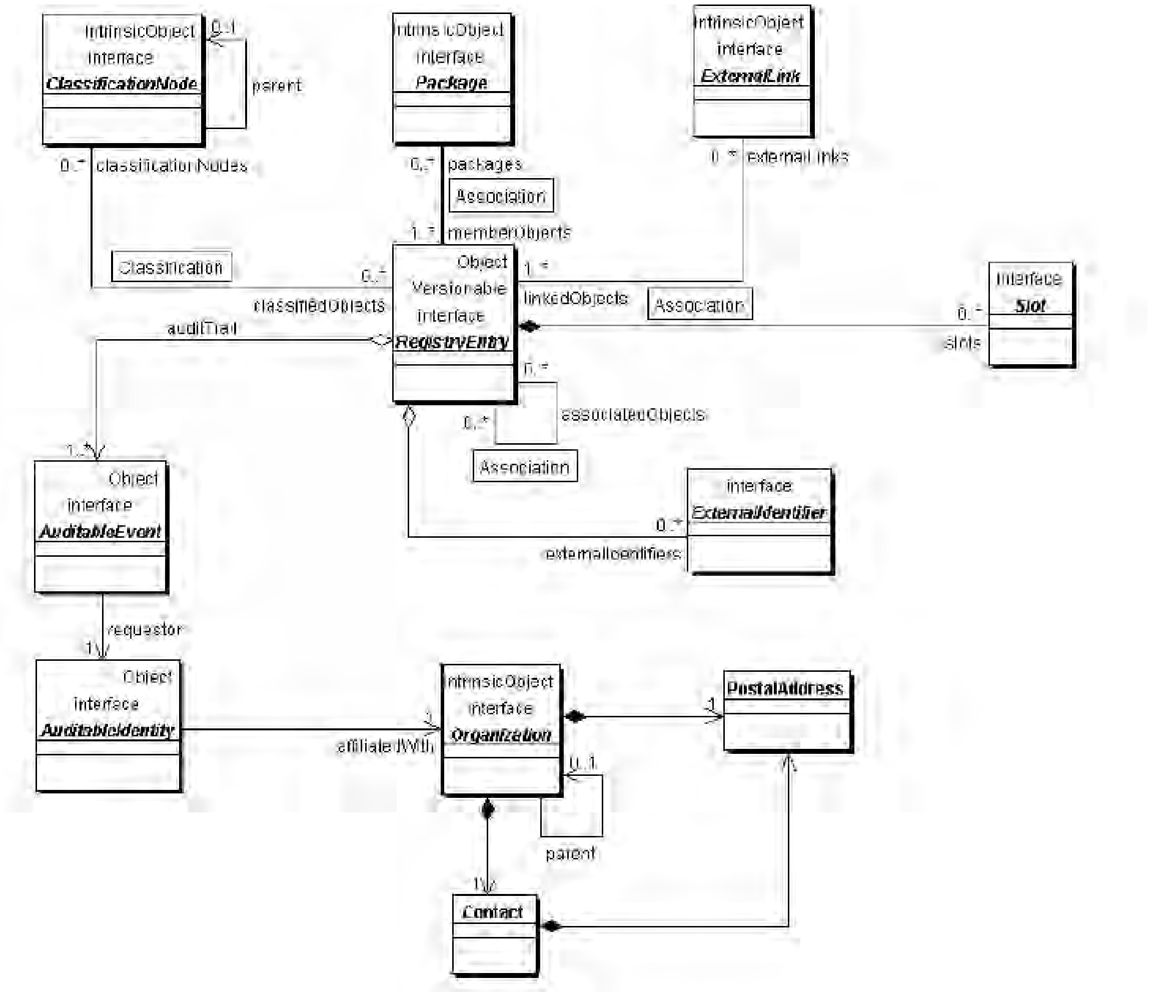
The ebXML process

The ebXML process is a joint initiative between OASIS (Organisation for the Advancement of Structured Information Systems) and CEFACT (Centre for the Facilitation of trade procedures for Administration, Commerce and Transport). The membership of the OASIS organisation is principally the software industry and the membership of the CEFACT working groups is principally user organisations. The ebXML organisation is therefore a blending of user knowledge and software development experience. The goal is to create a technical infrastructure that will facilitate B2B e-commerce, specifically in discovering who is trading electronically and what (XML) documents they use, and for registering XML specifications used in B2B e-commerce, based on the ebXML core component library.

ebXML is using UML as the modelling method and XML as the specification of ebXML service interfaces.

ebXML Registry model

UML model of the ebXML registry



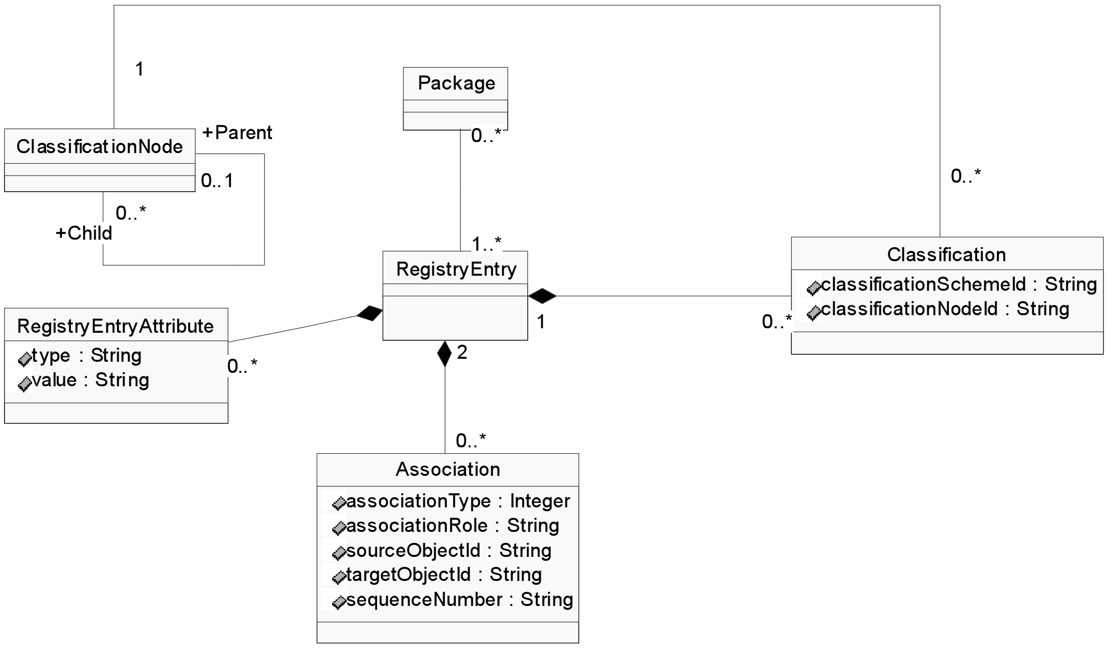
It is not the intention of this paper to describe this model. A reference to where the specification can be found is given at the end of the paper. The specification defines the methods for each of the classes, and the XML specification that implements the methods.

The main point about this model is that it can be used to store and retrieve objects according to any data model. For ebXML, the registry is used to access information about organisations, the products they buy and sell, and the XML documents they use for the commercial cycle (e.g. order, invoice).

IQML repository

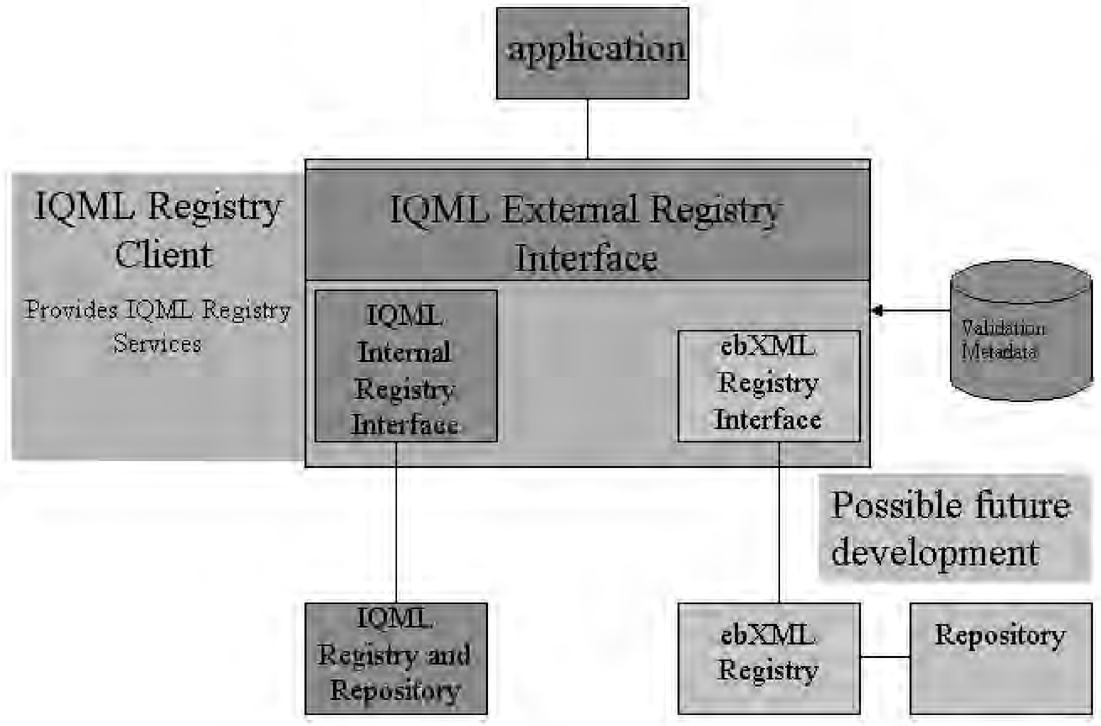
The repository for the IQML project is based on this model.

IQML Registry and Repository Model



The Registry Client acts as the bridge between the Applications and the Registry/Repository, and the intention is to be able to use an ebXML compliant Registry in the future, so that the underlying registry can be any registry that conforms to the interfaces specified by ebXML. The storage for the IQML Registry is XML.

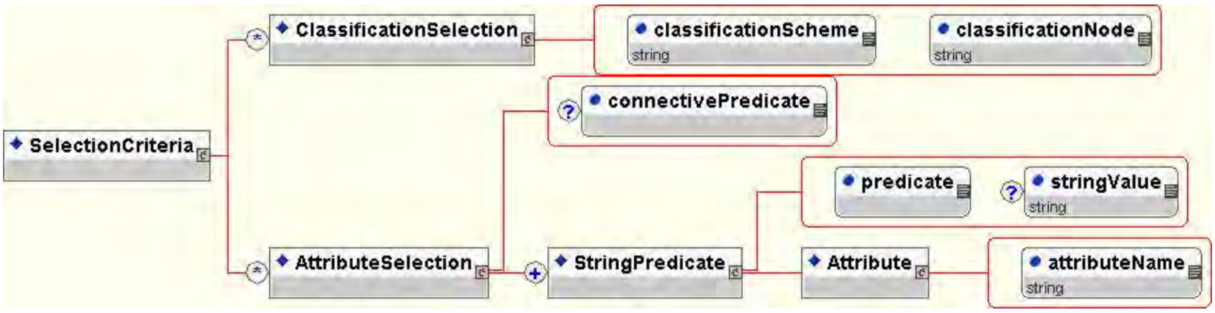
IQML Registry Architecture


application
IQML Registry Client
Provides IQML Registry Services
IQML External Registry Interface
IQML Internal Registry Interface
ebXML Registry Interface
Validation Metadata
Possible future development
IQML Registry and Repository
ebXML Registry
Repository

For IQML, we have based the Registry interface on a simplified version of the ebXML interface, but have borrowed many of the concepts from ebXML.

The interesting thing here is that these interfaces are themselves XML based. The following diagram shows the structure of the selection criteria for a query on the IQML registry.

Extract from the IQML Registry Query Interface



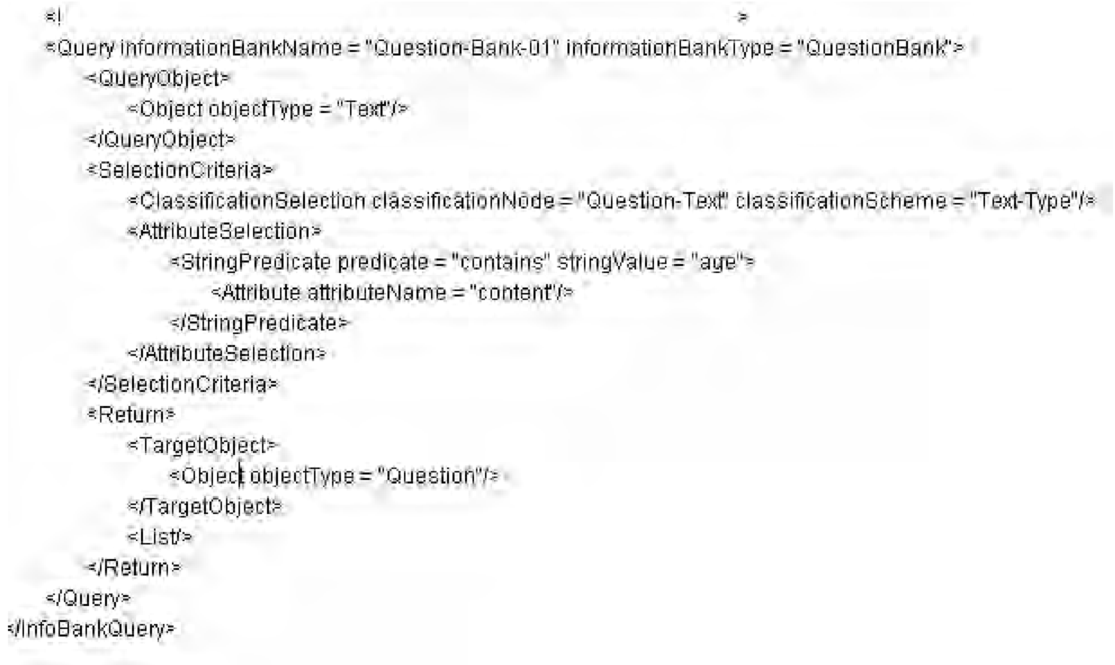
An Application, such as the Questionnaire Designer, will use the query mechanism. An example of an XML "message" that can be sent between the Questionnaire Designer and the Registry Client is shown below.

If you look at the extract model from the IQML Question Bank, then you will see that you should be able to make the following query of the repository.

> "Get a list of Questions where the question text contains the characters <age>".

The way this is done, using the XML interface is as follows:

IQML Repository Query in XML



```
<!                                                          >
<Query informationBankName = "Question-Bank-01" informationBankType = "QuestionBank">
    <QueryObject>
        <Object objectType = "Text"/>
    </QueryObject>
    <SelectionCriteria>
        <ClassificationSelection classificationNode = "Question-Text" classificationScheme = "Text-Type"/>
        <AttributeSelection>
            <StringPredicate predicate = "contains" stringValue = "age">
                <Attribute attributeName = "content"/>
            </StringPredicate>
        </AttributeSelection>
    </SelectionCriteria>
    <Return>
        <TargetObject>
            <Object objectType = "Question"/>
        </TargetObject>
        <List/>
    </Return>
</Query>
</InfoBankQuery>
```

It is interesting to note that XML is becoming the standard for integrating components that are loosely coupled. There is no need to implement CORBA (Common Object Request Broker Architecture –

OMG standard) or DCOM (Distributed Component Object Model – Microsoft), or Java RIM (remote invocation method) for object component integration, even remote components. Of course, these technologies are not obsolete and have an important role to play where their functionality is required (e.g. high throughput, high security, transaction based, tight coupling). However, where these functionalities are not so important, then XML is a valid and much simpler alternative. There are many XML tools that will help and many vendors are implementing XML as the way of integrating their tools.

It is interesting to note that there is an initiative under the Java Community Programme to develop a java based API for XML registries. The following is an extract from the relevant JCP web page (at java.sun.com).

Java™ API for XML Registries 1.0 (JAXR)

> JAXR provides an API for a set of distributed Registry Services that enables business-to-business integration between business enterprises, using the protocols being defined by ebXML.org, Oasis, ISO 11179

IQML use of OMG standards

The IQML Repository will also implement part of the CWM specification. The CWM is a metamodel and specification for the interchange of warehouse metadata. Within the CWM there is a relational package which defines the metadata of a relational database in a vendor neutral way. It is possible, using the CWM, to define the tables and columns of a relational database. Leading database vendors such as IBM and Oracle can read the CWM metadata and create a database. In IQML we will output CWM compliant metadata to define a database that can be used to store the responses from the questionnaire.

The CWM specification is, in fact, an XML specification of the CWM UML model using XMI.

XMI

XMI is an acronym for XML for Metadata Interchange. The original intent of XMI was for the exchange of UML models. The UML standard did not specify a syntax for the exchange of UML models. As more and more UML tools became available there was a necessity to specify a way that the model could be exchanged between the tools. So, XMI was born and it became an OMG standard in 1998.

However, the authors of the XMI specification knew that XMI was more powerful than its use for shipping UML models. It could also be used to generate an XML DTD from a UML model, and a document instance conforming to the DTD could be sent in XMI.

Java class libraries are available for reading and writing classes to and from XMI.

The Repository will use XMI to hold the specification of the underlying model for use by the Registry Client. The IQML Registry Client receives XML requests from an Application (e.g. Questionnaire

Designer). As the Registry/Repository is quite generic and allows any object to be associated with any other object, and allows an object a many attributes, the Client must have a way of knowing what is valid. In the initial XML specification, some of this knowledge was put into the XML schema (e.g. an enumerated list of valid object types). However, this has now been removed and replaced by an XML file of "Validation Metadata". This file describes the valid object types, attributes and associations. In the initial release of the Client this file will be quite simple, and will not contain all of the metadata required to do a full validation. However, it is intended to replace this file with an XMI file. The XMI will allow full semantic validation according to the UML model. This will give a totally "soft" Client, which can be used to access a repository based on any model. UML modelling tools can output XMI and it is this output file that will be read by the Registry Client.

Input to the standards process

CWM

The IQML project is also active in the standards making process. In this way it hopes to influence the standards emerging for XML based questionnaires, and to evaluate parts of the CWM. It is the raw data reporting that will feed the data warehouses and there is an extension metamodel in the CWM that is intended for this purpose – The Information Set. The IQML project will evaluate the Information Set as part of its research.

MDA

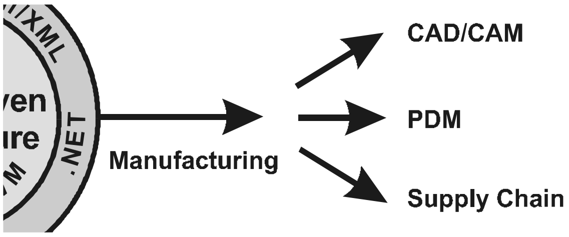
The OMG has announced recently a move towards the Model Driven Architecture (MDA). This is the most interesting aspect of the OMG work to survey computing. An extract of the MDA paper from the OMG is given below:

"Thus, in order to maximize the utility and impact of OMG domain facility specifications in the MDA, they will be in the form of normative, platform-independent UML models augmented by normative, platform-specific UML models and interface definitions for at least one target platform. The common basis on MDA will promote partial generation of implementation code as well, but implementation code of course will not be standardized.

.....

As another example, the Manufacturing DTF could produce normative MDA UML models, IDL interfaces, Java interfaces, XML DTDs, etc. for (Figure 2) CAD/CAM interoperability, PDM (Product Data Management), and a Supply Chain integration facility. Once the MDA models for these have been completed and adopted, their implementation can be partially automated in any middleware platform supported by the MDA.

Figure 2: UML models of frameworks for vertical facilities



The three facilities that we're using for this example – CAD/CAM, PDM, and Supply Chain – demonstrate a benefit that only the MDA can provide: CAD/CAM and PDM applications are tightly integrated and so will probably be implemented by an individual enterprise or software vendor in, for example, CORBA or EJB. Supply Chain Integration, on the other hand, is more of an inter-enterprise facility so we might expect an XML/SOAP based implementation supported by an industry market-maker or trade organization to become popular. But, it will be essential to interoperate among the three: CAD/CAM designs feed into PDM production facilities which drive the supply chain; in turn, the supply chain will refer back to CAD/CAM for details on a particular part. By starting all three out as UML models in the MDA, we may eventually be able to generate a significant portion of the implementation of each on its preferred platform, as well as the bridges we need to integrate each of the facilities with the other two."

The IQML project hope that there is sufficient interest from the software vendors and users to be able to standardise a UML for the questionnaire. This does not mean that there will be just one XML specification, but it does mean that the underlying concepts and semantic of the specifications will be consistent and therefore interoperable.

4. Conclusions from the standards process

For any domain that requires interoperability between software components from different software vendors, the way forward is becoming clearer.

- use the UML as the modelling tool
- standardise the UML model in a recognised standards organisation (e.g. OMG, CEFACT, ISO).
- develop syntax specific specifications from the UML model to suit the processing or exchange environments (e.g. XML, EJB, CORBA)
- Fine grained object interfaces will use the interface mechanism provided by the language (e.g. Java, C++)
- Coarse grained objects can be glued together using XML, especially when communicating over the web
- Object languages such as Java will see standards developed for interfacing to XML standards, such as XMI and ebXML specifications
- take advantage of international standards where these exist, and build on them

For IQML we have gained a lot from following standards:

- Use of common (UML) patterns in the development of the "user" model, especially in the area of abstraction
- Understanding of the flexibility of a generic registry/repository
- Good practice for the use of XML for loosely coupled components over the Internet
- Ability to take advantage of free software components that aid integration

We also hope to put something back into the standards process, so that others can gain from our experience. Users (rightly) expect a lot from software products and it is only by using standards that the software industry can deliver the requirement. Car owners do not expect to have to buy petrol from only one oil company, or windscreen wiper blades, or sparking plugs. Software components must be inter-operable and this means adopting standards. The software industry is very active in the making of these standards, and we can all gain from this.

References and Acknowledgements

ebXML Registry Services Specification: http://www.ebxml.org

ebXML Registry Information Model (RIM) http://www.ebxml.org

Model Driven Architecture by Richard Soley and the OMG Staff Strategy Group: Object Management Group White Paper Draft 3.2 – November 27, 2000. Extract reproduced by permission of the authors: http://www.omg.org

UML, CWM, CORBA, MDA, XMI are trademarks of the OMG. OMG logos reproduced by permission of the OMG.

ebXML material is the copyright of ebXML

IQML material is the copyright of the IQML project and is reproduced with permission of the IQML project

About the Author

Chris Nelson has been working on EDI standards issues with Eurostat for 10 years. He played a key role in the development of a model and EDIFACT specification for a message to exchange multi-dimensional data and time series. This message is used by EU/EEA statistical organisations, including Eurostat, and the ECB and National Central Banks. He is chair of the Analytical Data Management Special Interest Group of the OMG. He has been involved in the ebXML process, particularly on the Registry and Repository, and the Core Components groups. His company, Dimension EDI, is a co-submitter of the OMG Common Warehouse Metamodel (CWM) standard, and is involved in the IQML project.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

Metadata, Semantic Interfaces and Meaning in Global Communication

Reinhard Karge

Abstract

The paper discusses features of metadata-based communication related to statistical[9] data. It shows the way metadata can be used in communication processes and the problems resulting from this strategy. It demonstrates also, how terminology models and semantic interfaces can be used to overcome the problems and to establish worldwide communication based on statistical metadata.

Keywords

Terminology model, metadata, information exchange, semantic interface, XML, metadata database

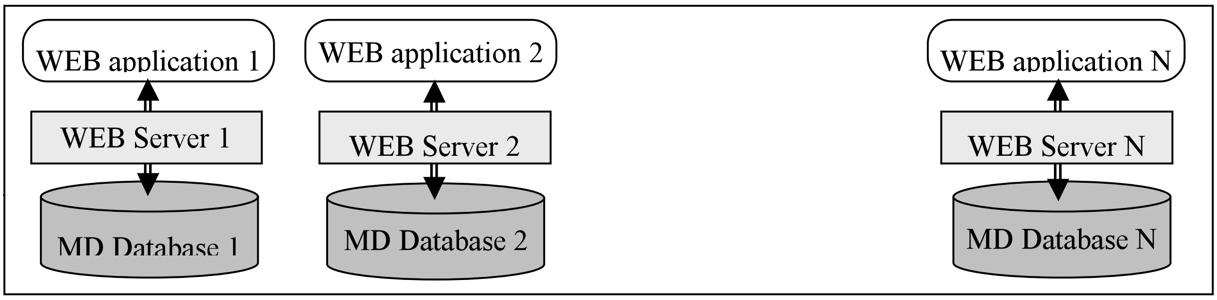
1. Introduction

Metadata is an essential part of statistical information. In most cases we cannot find, process or understand data without metadata. According to these three purposes metadata can be characterized as

- Physical metadata
- Operational metadata
- Conceptual metadata

There might be other aspects but in this paper we will focus to these ones. The three aspects of metadata mentioned are required when we refer to statistical data. We need physical metadata to locate statistical data. Otherwise we could not retrieve it. We need conceptual metadata to understand the meaning of data and we need operational metadata to see the process that has produced the data.

But how can we get statistical data and metadata from different statistical offices? Since all statistical offices are organized in different ways and are using different databases for data and metadata, an ex-



ternal user has to learn how to get statistical information differently for each one. Comparability of data is also difficult to achieve and needs additional consultation with experts.

This works sufficiently as long as only data is retrieved and the user has some knowledge about statistical concepts. But if one wants to get more detailed information about the data and the information available, then metadata-based retrieval functions become necessary.

Statistical metadata objects carry different aspects of metadata. Conceptually we do not make any distinctions about whether the metadata associated with a certain metadata object is physical, operational or conceptual. During the last two years different terminology groups [6] have described about 70 statistical metadata objects belonging to four different areas:

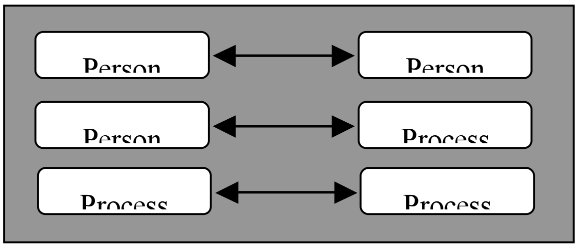
- Classification
- Variables
- Registers and tables
- Thesaurus

We can find those statistical metadata objects in all statistical offices, but they might be named differently. Statistical concepts are expressed by means of statistical metadata objects, i.e. by giving a name to the concept and describing its attributes and relationships to other concepts. To understand each other we have to agree on names for the concepts or metadata objects.

If a requirement for human communication is a terminology agreement this is even more valid for communication on a more technical level. Whatever format we choose, XML, GESMES or CLASSET – we need a common understanding of the "terminology" used in the document or file containing the data. Technically terminology is expressed mainly in data models. Unfortunately, the concepts and details are not named according to the terminology in human language. The attempt to agree on common metadata database models has not succeeded so far. Hence, we usually cannot understand a database model without translation.

2. Terminology standards for communication processes

Communication without metadata is impossible or senseless. In the past operational and physical metadata has been used in production processes. Now we are going to make available more and more conceptual metadata. This supports the production process on the one side but it is also the base for understanding the statistical data. Some conceptual metadata is contained in statistical tables in the form of headings for rows and columns, footnotes, or explanatory text. This is typical metadata provided for "Person to Process" communication. But metadata is required between all types of communication. And in all cases, communication requires an interface, an agreement that allows the exchange of information between objects. We consider here only persons and processes as communicating objects, even though there exist many other communicating objects. With this restriction we can distinguish three types of communication: "Person to Person", "Person to Process" and "Process to Process" communication.



"Process to Process" communication is not well developed from this aspect, though we do have agreed syntaxes, such as C, SQL and XML. In contrast to natural language, there is no general agreement on semantics in process languages. Programmers choose names for program variables or database attributes according to their taste. When reading an SQL statement or a C-program it is in general not possible to interpret the meaning of the syntax because the intention of the specific technical names (terms) is not specified. This causes problems not only in "Person to Process" communication but also in "Process to Process" communication. You cannot just run the same SQL query on Swedish census data and on US census data, even though these databases may display the same variables. Hence, a semantic standard is required that is based on common terminology.

Based on common terminology models a semantic interface will map the agreed terminology to the technical database model. This allows communication on the user and application level based on the common terminology model. Because communication designed this way transfers exactly what the user has defined it will be consistent with the user's requirements. The user can express exactly what he has defined in his terminology model and it is left to the implementation of the semantic interface how far the user's requirements can be supported or not.

Communication standards have to fulfil the requirements of passive and active communication.

Passive Communication

Passive communication is more like sending a parcel that contains all that is needed. Passive communication is typically for "person to person" communication. But it is used also in "process to process" communication if active communication is not possible. Passive communication is based on files that are created by one communication partner and send to the other process partner. Passive communication is always based on a medium (file) that contains the information to be transferred.

For passive communication XML is becoming a syntactical standard. Since XML allows the creation of well-structured documents based on a given XML schema (as well as simple structured documents) it seems to fulfil all requirements for passive communication. This allows the exchange of statistical tables with metadata as well as classifications, variable definitions or simple documents. Moreover, XML provides extensive presentation features (using XML style sheets) that allow transferred data to be presented in an appropriate way.

Active Communication

Active communication is more like a telephone call, which consists of a sequence of questions and answers. In this case the content of the communication is not pre-determined but depends on the specific requirements of the dialogue partners. Hence, active communication is typical for maintenance or retrieval processes. Active communication is required when the content of the communication is not pre-determined. This is the case e.g. when creating or maintaining metadata or when defining process requirements e.g. for a table.

There is no unique standard for active communication but COM and CORBA can be considered as such technical standards. Considering COM[10] as an agreed technical standard we need further specifications to be able to communicate in different environments. A number of communication objects and

[10] Since CORBA and COM are quite similar it does not make any difference to consider only here COM.

interface functions must be defined as a type of COM protocol that enables different parties to communicate with each other. The common metadata interface ComeIn [8] is a possible specification for a generic metadata interface that allows access to any type of metadata object, independent of the metadata database where it is stored. But even this is not sufficient. Even if we know that there is a function that creates a metadata object server (CreateMOS) we still need to know the name of the object to be referenced. A call such as CreateMOS("KeySystem") this will only work in an environment where "Key System" is known as a metadata object type. Thus, we have to add to the technical interface specification a semantic specification that defines the names for the objects and attributes to be accessed. Doing this we will be able to provide active communication that works in any environment that supports the agreed standard.

Conclusions

1. Communication between persons and processes requires a terminology agreement (terminology model).
2. Terminology models must support active and passive communication.

3. Defining terminology models and semantic interfaces

Statistical offices and other organisations are not only asked for data but, increasingly, metadata is requested to support interpretation of data. Statistical metadata can be considered as a knowledge base for a statistical office. It describes processes as well as concepts and meaning of statistical data. Transferring knowledge about statistics means transferring metadata.

Example: When getting data for income by age groups, sex and education we need not only information about the data structure (physical metadata), but also about currency and precision (operational). Moreover, we need information about the concept of the income variable (conceptual metadata) that describes what the income includes.

To be able to interpret statistical data correctly we need conceptual metadata in addition to operational and physical metadata. While statistical data is transferred in the form of numbers, which are language independent, statistical knowledge is expressed in human language. To be able to understand each other we must speak a common language. Even when we agree upon one communication language (e.g. English) we will still have problems as long as we are using different terms for similar concepts. New formats for data exchange such as XML allow transferring metadata together with data. But we can only understand the metadata passed when it is based on a common terminology.

Since 1999, several terminology groups have started to define common terminology models for statistical concepts. METANET is an initiative from Eurostat that tries to harmonize different metadata models. During the same time technologies have been developed to integrate terminology models into software development.

There is quite deep knowledge about physical and operational metadata available in different statistical offices. These aspects of metadata are well formalized and used mainly for production purposes. Conceptual metadata was not formalized in such a strong way. But there are also attempts to describe formal roles of providing conceptual metadata, such as in SCBDOC [1], the Swedish system for documenting statistical activities (products and surveys). Even though those documentation rules give a quite formalized definition of conceptual metadata they are not sufficient for providing metadata related to a specific piece of information, e.g. the definition for a specific variable. One can, of course, go to the document and try to find the variable's definition in the document but this is quite complicated and unsafe. While the SCBDOC is very specific the Dublin Core [9] defines very general rules for providing conceptual metadata for any type of data item by means of metadata. The Dublin Core specification allows to describe anything as an "element" but is does not provide rules for describing specific metadata objects as classifications and others.

The IMIM project of the 4th Framework has defined a strategy to provide metadata on the level of atomic metadata objects. The Bridge model [2] shows a way to define metadata based on well-defined metadata objects on different levels. As a technical model the Bridge model is too difficult to understand for non-technical persons and it is just another way of storing and providing metadata. But in contrast to other metadata systems Bridge includes the three aspects of metadata. It is possible to provide metadata for a particular metadata object, such as a variable in a table, as well as documentation for more general metadata objects. But just like other metadata systems it has its own (technical) terminology, which is hard to understand for persons not working on that level.

Defining terminology models

During the last two years several groups have been working on defining conceptual models for statistical metadata based on terminology definitions (terminology models). The terminology model for classifications and related objects (Neuchatel group [3]) was finished in March 2001. Terminology models for variables, registers and cubes and the thesaurus are under discussion and are available as first drafts [4] [5].

Terminology models can be defined by defining names for statistical concepts such as **classification** or **classification item** and its attributes and relationships to other concepts (characteristics or details). Such two level terminology definition correspond directly to a conceptual metadata model were the concepts are considered as metadata object types and the details as properties (attributes and relationships) of these object types.

The meta-model for a terminology model is simple and easy to understand for non-technical persons. The terminology model consists of two object types:

Concept The concept defines a term that refers to a basic idea in a subject matter area. The names for concepts and synonyms must be unique in a terminology model. The concept is defined by:

Name The name is a single word or group of words that identifies the concept.

Example: *Classification Version*

Description A description or definition of the named concept

Characteristics List of characteristics that describe the details of a concept. Characteristics are defined as Characteristic.

Synonyms List of synonyms that can be used instead of the concept name.

Example One or more examples that illustrate the concept

Characteristic The characteristic of a concept defines a relevant detail (attribute) of a concept. The names and synonyms for characteristics must be unique within a concept definition. The characteristic is defined by:

Name Single word or group of words that identifies the characteristic

Example*: title, items* or *registration authority* for a classification version

Description The description or definition of the named characteristic.

Concept If the characteristic is not simply defined as text but refers to another concept the referenced concept is mentioned here. In documents referenced concepts are visualized as underlined terms.

Example**:** organisation that owns a classification

Synonyms List of synonyms that can be used for the characteristic.

Example One or more examples that illustrate the characteristic

In contrast to other terminology definitions terminology models differ between context independent terms (concepts) and concept related terms (characteristic). Thus, context related terms might be defined with different meanings in different contexts or for different concepts. The terminology model should also include rule definitions for defining rules for concepts and characteristics. This has not been included in the terminology meta-model so far to keep it as simple as possible.

The following example is part of the Neuchatel group terminology model for classifications. It demonstrates one way of defining a terminology model based on the rules described above:

Classification A classification defines the general idea of classifying statistical *observation units* in a given population. A classification is a structured list of mutually exclusive categories to describe all units of a defined population according to a defined property.

A classification is represented by one or a number of consecutive classification versions which may be linked over time by correspondence tables.

Name A classification is identified by a unique name, which may typically be an abbreviation of its title.

Title A classification has a title as provided by the owner. The title usually indicates the scope of the classification.

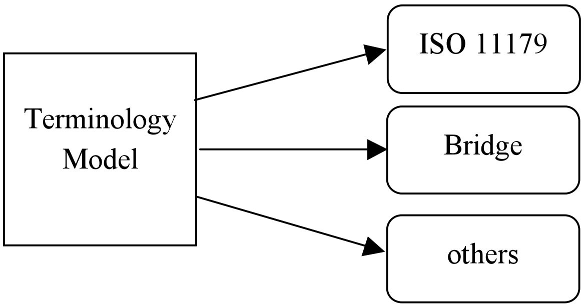
Owners The statistical office or other authority (organisation), which created and maintains the classification. A classification may have several owners.

Units classified The classification may refer to a number of statistical objects.

The meta-model for the terminology model is a recursive model, i.e. it can be described as a terminology model again. Thus, the terminology meta-model is an example for a terminology model as well.

Each conceptual metadata object may carry all aspects of metadata. Our experience has shown that there is no problem in agreeing on a common set of concepts and their characteristics.

From terminology model to semantic interface



Terminology models provide a good explanation of concepts in a certain subject area. Thus, terminology models become a good base for communication as soon as we can agree on a certain terminology model. Unfortunately, terminology models are not consistent with metadata models in different environments. Nevertheless, there is a relationship between metadata objects (relations) and its properties (attributes) in a metadata model and concepts and characteristics in the terminology model. The mapping from metadata models to an agreed terminology model can provide a common communication base. Communication based on a common terminology model allows exchanging metadata between different metadata databases.

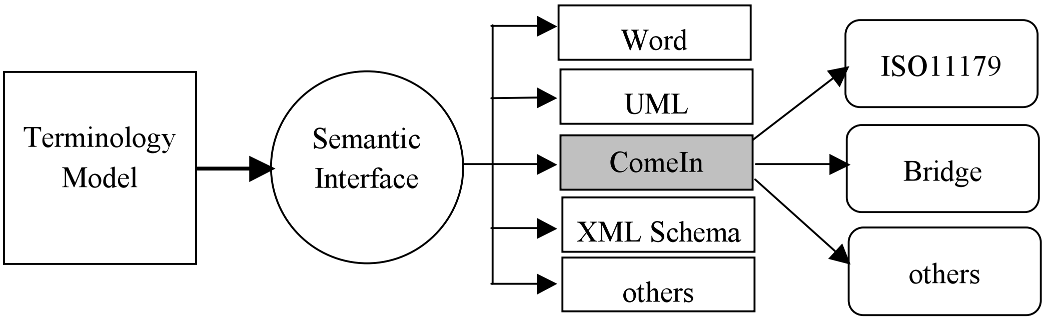
Unfortunately, terminology models cannot be used as such for technical communication because names of concepts and characteristics, which correspond to object types and properties, might consist of several words. Hence, the terminology model must be transferred to a semantic interface.

A specification that includes a technical and a semantic specification is called a **semantic interface**[7]. A semantic interface specification might result in different implementations, such as XML schema or ComeIn. Communication based on a semantic interface becomes database and database model independent because the semantic interface maps the terminology model to a specific metadata model.

Semantic interfaces can easily be constructed based on terminology models. Since terminology models are database independent the semantic interface becomes database model independent and can be used in any environment. Using the same semantics for the XML schema and the semantic interface supports both importing or exporting statistical metadata directly via the semantic interface.

The semantic interface maps concepts and characteristics from the terminology model to logical object types and attributes, which follow technical naming conventions. Moreover, some technical information is added to the semantic interface describing the cardinality for references, inverse references for relationships and sort keys for collections. In some cases additional objects and properties, which are not part of the terminology model, must be defined in the semantic interface as well to map to undefined elements in different metadata models. In this case the terminology model is expanded by “internally” defined concepts and characteristics.

The definition of a semantic interface is the base for many different purposes. It can be used for generating documentation (word document or UML), for generating technical interfaces that map the logical object types and properties to database object types (relations) and their properties (attributes), for generating XML schema for metadata data exchange, for comparing (mapping) metadata models and many other purposes.



Even though the technical problems seem to be solved in principle, coordination work is necessary to organize the building and creation of common terminology models.

Conclusions:

1. Technically it is possible to store and retrieve all aspects of statistical metadata in/from a metadata base.

2. Terminology models are a possible way to describe the concepts in a specific subject area and can be used as common communication base for human communication as well as for technical communication.
3. To make metadata available for different types of users and processes a semantic interface is required that maps to different metadata models.

4. Using semantic interfaces

This chapter gives some examples for using semantic interfaces for different purposes. Especially when communicating between different organisations interface agreements are required that make communication possible. A semantic interface based on a terminology model is a way for defining a communication standard between organisations on conceptual level.

Metadata exchange

Metadata exchange is one of the requirements for data providers. Classifications or variable definitions are requested in addition to tables as well as quality metadata. New formats for data exchange as XML allow transferring metadata together with data. Nevertheless, XML is just a syntactical standard and will not solve semantic problems. The following example shows a simple classification in XML format:

```
<KeySystem> <name> Sex </name>
            <key>      <code> 1 </code> <text> male </text>      </key>
            <key>      < code> 2 </code> <text> female </text>   </key>
 </KeySystem>
```

This example refers to a proprietary terminology. The consequence is that the XML document cannot be processed without being “translated” to the terminology used on the recipient side.

Without a common XML schema N*(N-1) transformations must be provided for N different XML schemas. However, with a common XML schema the number of transformations is reduced to 2*N (from and to the common XML schema).

Several attempts to define a common metadata model have failed. Semantic interfaces, however, seem to be a possible way to agree on common standards. An XML schema can be derived directly from a semantic interface.

```
<ClassificationVersiion> <identifier> Sex </identifier >
      <items>
            <ClassificationItem> <code> 1 </code> <title> male </title>
            </ClassificationItem>
            <ClassificationItem > < code> 2 </code> <title> female </title>
            </ClassificationItem>
      </items>
</ClassificationVersiion>
```

This example looks similar to the first one except that it is using different names. However, it is referring to common terminology agreements and can be easily exchanged between organisations, which accept these agreements.

Since a terminology model expresses exactly the conceptual level of statistical knowledge, XML based on a common terminology model will always be able to carry all the required information. Agreeing on common terminology models will open the future for providing common enhanced presentation roles for statistical data and metadata and other XML based features.

The semantic interface is a good base for defining an XML schema. An XML schema is a proposed XML standard that defines the structure of an XML document. It is comparable with DTD (Document Type Definition) but especially developed for data exchange. The semantic interface allows generating XML schema nearly completely from the terminology model.

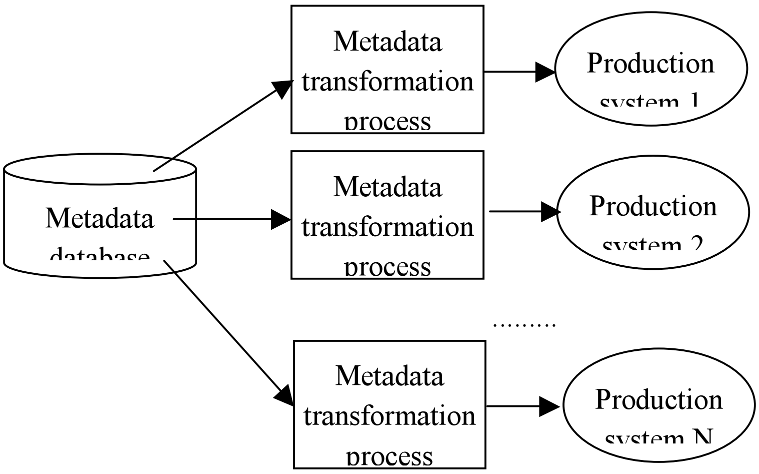
Conclusions

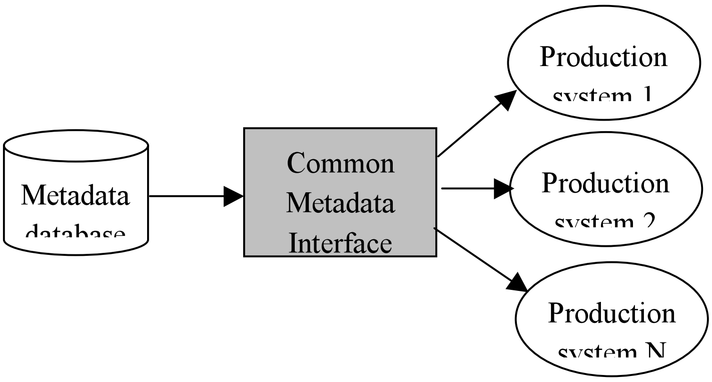
1. A semantic interface based on a common terminology model is a base for exchanging metadata between different organisations.

Metadata in production systems

One big challenge is to increase the usage of conceptual metadata in statistical production systems. So far metadata is provided for different processes in the appropriate format that is requested from the production system. That means that a number of metadata transformations have to be provided in each organisation to get the production system running.

This makes it always difficult to get a production system running in a specific environment. A semantic interface is a solution for these problems as well. When combining the semantic interface with a technical implementation as ComeIn, processes can use either an active interface or standard mappings between the semantic interface and the production system can be used.


Metadata database
Metadata transformation process
Metadata transformation process
Metadata transformation process
Production system 1
Production system 2
....
Production system N



ComeIn tools and SuperSTAR are two examples for commercial software products that provide metadata access by referring to an active interface. Providing metadata transfer from ComeIn to the production system can support other production systems that are not ComeIn compliant. This has to be implemented once and not separately for each organisation. The only step that is required for each organisation is to provide the mapping between their metadata databases and the agreed semantic interface.

During the last year the common metadata interface ComeIn has been developed and is available as a public domain specification. Based on terminology models ComeIn has been provided as a semantic interface that supports all common terminology models known so far. Even though software companies are highly motivated to support semantic interfaces, they cannot define appropriate terminology models. Only experts in the organisations can do this.

There are already several standards that could be considered as semantic interface specifications (ISO 11179, Dublin Core, DDI). However, these standards are too general and do not allow transferring the requested metadata for e.g. classifications or variable definitions to the production system.

Conclusions

1. Metadata references on a technical level should be based on the common terminology model.
2. Tools and applications should communicate with metadata via a semantic interface based on the common terminology model.
3. Integrating metadata in the whole production process (including conceptual metadata) will allow a high degree of automation of statistical processes.

Online Metadata

Modern presentation of statistical data will include statistical metadata for the user, giving enhanced explanations of statistical data shown in a table, whether in the Internet or in an enhanced tabulation system. This requires metadata on the base of conceptual metadata objects. It is always hard to get this information directly from a Word document or other complex documentation. Different components of a statistical table, such as categories (lines or columns) or variables (columns) can, however, be linked with metadata stored for these concepts or metadata objects. Since the conceptual metadata model (terminology model) defines links to other concepts (metadata objects) one can follow those links to get more information related to the object of interest. This functionality is not limited to the concept of hyperlinks when presenting tables in the Internet. Various tabulation systems, such as SuperSTAR, are able to provide metadata together with the table and allow traversal of the links to other metadata objects.

That means that statistical tables have to be provided together with metadata. To guarantee consistency between statistical data and metadata, statistical metadata must be linked to the whole production process of statistical data. A safe way to get e.g. information as the question that has been asked to get the statistical data is to go to the questionnaire or to follow the link of the variable to its source in the questionnaire.

For this purpose modern production systems (as the tabulation system SuperCROSS) store a metadata hook for each metadata object that has been retrieved from the metadata database (for variable definitions, classification items and others). The metadata hook is a unique identifier for a metadata object in the database.

When requesting background information (e.g. by right clicking on a row or column label in a table) the system calls a hook server that shows some basic metadata information for the requested object (e.g. the definition of a data element ‘income’). Moreover, it provides links to other metadata objects that can be displayed. A hook browser can be provided as simple WEB application or as a specific component of the production system.

Conclusions

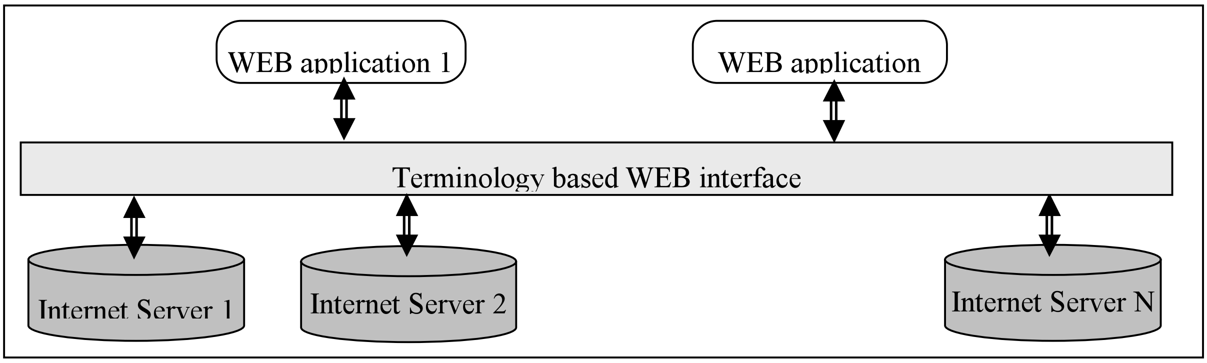
1. Supporting metadata hooks is the base for providing online meta-information systems.
2. General hook servers can be easily provided based on a semantic interface as ComeIn.

Metadata based retrieval functions

Retrieval functions usually require knowledge about the concepts the facts we are looking for are based on. When asking for something that is related to income one might get many hits and it becomes quite difficult to find what we are interested in. But if we know about the concept of variables we can ask for income related variables first. Then we can select some variables that reflect the information we are looking for. In the next step we can ask for the tables (assuming that we are aware about the concept of statistical tables) or cubes containing this variable. We can restrict the information to specific regions or age groups when we know the concept of classifications and value sets.

This shows very roughly how metadata based retrieval functions can work and how statistical metadata can be used to provide enhanced retrieval functions. But again we will have big problems when we search for information in different environments. Technical features such as XML can help to provide powerful applications but they will not solve the terminology problem. Metadata databases, which are filled with more and more knowledge about statistics, will be built in the near future. But to make worldwide use of this knowledge requires a common terminology that allows the right questions to be asked.

Common terminology models can solve this problem. In addition to metadata tools and production support, WEB applications can be built based on common terminology models. In this case statistical metadata is available in the same way independent on the office where it has been created and stored.



Using terminology based WEB interfaces allows us to combine keyword search and object link techniques in different WEB applications, regardless on the metadata storage used in a specific environment. WEB applications that are based on common terminology models are intuitive and easy to understand because they refer to the same terminology in each context.

A retrieval function is considered as a function that returns a number of WEB pages, or more generally, a number of (meta) objects. Retrieval functions based on terminology models are more flexible than simple keyword search functions, since we can combine three search techniques that in a WEB application:

- Keyword search technique

    Keyword search allows searching for all objects that contain a number of defined search words or that are linked to these terms. Keyword search is provided as free text search (slow) and index search (fast).

- Reduced scope technique

    This technique is usually combined with keyword search. By defining the type of objects the number of hits can be extremely reduced and the search process will be much faster. One typical way to reduce the scope is to search for objects of a given type (e.g. Variables or Tables referring to the search or key word).

- Object link technique

    Object link techniques are based on the fact that a number of relationships exist between different (meta) objects. Thus, it is quite common in many cases to follow the links between objects, which are part of the knowledge stored in the metadata database.

Combining keyword search facilities with linked object techniques and reduced scope features allows flexible, fast and user-friendly search strategies. XML can be used as well as Active Server Pages generating HTML direct from the metadata database.

Conclusions

1. XML and modern metadata databases allow the combination of three useful search techniques.
2. Retrieval functions become metadata database independent, when being based on common terminology models.

References

[1] User Handbook for the Documentation System SCBDOK with the Computer Support PCDOK

[2] Bridge – Object Architecture, IMIM work papers[11], 1998

[3] Classification terminology, Neuchatel group 2000, http://www.run-software.com/downloads/ClassificationTerminology.doc

[4] Variable terminology, Oslo group 2000, http://www.run-software.com/downloads/VariableTerminology.doc

[5] Terminology for statistical activities, cubes and registers, 2000, http://www.run-software.com/downloads/ActivityTerminology.doc

[6] ComeIn Overview, run-Software 2000, http://www.run-software.com/downloads/CIOverview.doc

[7] Semantic Interface, run-Software 1999, http://www.run-software.com/downloads/CISpecification.zip

[11] IMIM working papers are available at http://imim.scb.se

[8] ComeIn User's Guide, run-Software 2000, http://www.run-software.com/downloads/CIUsersGuide.doc

[9] Using Dublin Core, http://dublincore.org/documents/2000/07/16/usageguide/

About the Author

Reinhard Karge has been working on statistical metadata since 1980. He was involved in developing a statistical database system for the East German statistical office (1981-1989) and in developing the Base Operator System for processing relational data and metadata on a logical level (ECE project 1983-1986). The system provided basic relational operators that could work on different types of relational databases in a mainframe environment. Beyond data operation it includes appropriate metadata operations.

1990 he started to run a private company. There he lead the development of the object oriented database management systems and tools (ODABA2). ODABA2 is an OODBMS on a high conceptual level. It supports a number of features following principles of reflections in natural language. From 1997 to 1999 he lead the development of the integrated metadata system Bridge (IMIM project). The task was to build a highly integrated metadata system for statistical offices. The project finished in April 1999. The successor of the project, Bridge^{NA}, is an ODABA2 application and in use in statistics Sweden, Switzerland and Norway.

Since the IMIM project finished in 1999 he has been working in different terminology groups for defining common terminology models for different metadata areas in statistics and in developing the semantic interface ComeIn and a series of ComeIn tools for handling statistical metadata.

The idea behind all these activities is to bridge the gap between linguistic and database models. This will open a new world in IT technologies and can be considered as one step forward to artificial intelligence.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

On the Use of Metadata in Statistical Data Processing[12]

Jean-Pierre Kent

Abstract

Statistical metadata are data on statistical data. They are essential both to statisticians, who need them to control the process, and to end-users, who need them to obtain insight into the data. Experience shows that there are many threats to the quality of metadata. The relation between data and metadata may be disturbed, the metadata actually used may diverge from the initial intentions of the designer, and the metadata specified by design may diverge from the standards.

This paper describes a long-term view on how the statistical process should be set up in order to maximise the effectivity, coherence, and reliability of statistical metadata. It proposes to create metadata-aware tools to support metadata-driven processes. The statistical process is seen as a set of tasks to be performed under control of metadata.

Keywords

Data, Metadata, Meta-information, Statistics, Process, Concepts, Standards, Co-ordination

1. Introduction

Metadata satisfy two primary needs: they *guide* statisticians in the processing of data and they *inform* users about the meaning of the data. They are, therefore, indispensable. Indeed, statistical metadata have existed ever since statistics came into existence.

The last decade of the twentieth century saw a growing concern for metadata. Advanced information technologies, with highly automated processes and integrated dissemination systems as their main exponents, gave rise to the need for a more disciplined and thoughtful handling of metadata. Originally, the focus had been on supplying comprehensive and well-specified documentation to accompany the data (Sundgren, 1973, etc.). Such documentation is primarily aimed at making clear to end users *what* the data are about (statistical concepts) and *how* they have been produced (statistical processes). In this approach, there is no operational link to the metadata used to guide statisticians in charge of running the process.

[12] This is an abridged and updated version of a paper published in 2000 by Jean-Pierre Kent, Jelke Bethlehem, Ad Willeboordse and Wilfried Ypma.

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

Traditionally, therefore, metadata controlling the processes are directed primarily *towards* humans operating them, while metadata documenting the results are produced *by* humans. Such metadata are, therefore, primarily *human-oriented* and *human-based.*

This being so, two questions arise. First: can we be certain that the processes really behave as specified? Second: do the metadata published with the data indeed provide a full and truthful description of these data? "Metadata collection is dull, expensive and time-consuming", Sundgren sighs. Considering this implicit warning, one might indeed have doubts about the quality of metadata, especially with respect to keeping track of changes in processing conditions, methods, and concepts.

This paper describes the goals and problems of this shift in the focus of metadata. The essential element in our strategy can be stated as follows: maintain the tie between metadata and data throughout all stages of the statistical process, so that the metadata at every stage meet three demands:

- they provide a fully specified description of the data they accompany and of the processes affecting the data;
- they reflect the intentions of the statistical design;
- they comply with general standards, with respect to both statistical concepts and processing methodology.

Before elaborating on this, we have to make clear what exactly we mean by *metadata.* Indeed, although this notion seems to be self-explaining, discussions in the past have shown that the concept, if not properly defined, may lead to confusion. Defining metadata will be the subject of section 2. In section 3 we will expose the three fundamental goals of our metadata policy. We will then present a model of the statistical process (4) within which to implement our goals (5). Section 6 will explore implementation aspects more in depth. In section 7 we will distinguish the roles of humans and machines in the statistical process. Finally we will present in section 8 a view of what has been achieved at SN in the light of the views presented in this paper.

2. What are metadata?

The simplest, but most widely accepted definition tells us that *metadata* are data about data, like *metaphilosophy* is philosophy about philosophy. Therefore, metadata are data. Not just data, but a specific sort of data.

Specific definitions, however, cover widely different concepts. In Olensky (1996) we read: "Metadata are information standards." In the *Blaise III Developer's Guide* (1994, p. 51) we read: "[Metadata] is a formal statement of the information used for collecting, processing and publishing survey data." There seems to be very little overlap between these two definitions. While Olensky stresses standardisation, the Blaise manual points to formalisation. Olensky refers to information in general, whereas the manual mentions only the metadata used for carrying out a survey. Most definitions fall between these two extremes.

It is a matter of experience that communicating on the subject of metadata is not an easy task. Because the concept covers all aspects of data processing, specialists with very different backgrounds, aims, and expectations approach it from different points of view. Considerable effort has to be dedicated to understanding each other, even within the same institution.

A general definition of metadata

In order to cover all concepts commonly referred to as metadata, one can define statistical metadata as:

> All the information needed for and relevant to collecting, processing, disseminating, accessing, understanding, and using statistical data.

Metadata are relevant in the following areas – see Kent & Schuerhoff (1997):

- definition of statistical concepts;
- modelling of data and processes;
- storage structures and transfer protocols;
- standards to ensure a uniform and co-ordinated approach;
- information about availability, location, meaning, quality and use of data.

Such information is needed in order to support good communication between all parties involved with data. These parties are:

- *data providers* and *interviewers*, who need to understand what data have to be collected;
- *statisticians*, who are responsible both for designing and running the statistical process;
- *end-users*, who have to be able to find, select, understand, evaluate, and analyse the data for their own purposes.

In order to support and automate these activities, software also needs this information.

Metadata can appear in different shapes, depending on the goal they are used for. A *questionnaire*, for example, describes data in terms of questions about data to be collected. It enables respondents to supply the data in the form of answers. *Documentation* is used to inform about the structure and meaning of the data and/or about the logic of a process producing data. *Titles, labels* and *explanation texts* describe the variables involved in a table. A *database schema* or a *record description* supplies software with the information required for automatic access to the data.

None of these manifestations of metadata gives an integral picture. All of them offer a *selection* of information relevant for a certain *goal* in a certain *context*, and present it in a *format* appropriate for meeting this goal. Therefore, the different manifestations of metadata are not suitable for comparison and integration. What is needed is a new way of representing metadata. We need to be able to express all relevant information about the data in an abstract way, with reference to meaning and structure, rather than to selections, display formats, goals or contexts of use. This abstract representation can then be used to generate selective views of the metadata suited for specific use in specific contexts.

Aspects of metadata

Metadata, as defined above, apply at different moments and places in the statistical process, where they may fulfil different roles. It is important to be aware of the different qualities in which metadata occur. In the following, we provide some distinctions that are relevant in the context of our policy.

Metadata refer either to *states* of affairs or to *processes*. State metadata can describe a data set, while process metadata will describe how one data set is transformed into another. While state metadata usually define statistical concepts (like the variable "turnover" or the classification "economic activities"), process metadata primarily refer to methods (like imputation or estimation methods). Therefore, our distinction between *state metadata* and *process metadata* reflects the dualism of *concepts* vs. *methods*.

This distinction is made at the level of statistical concepts, and should not be confused with the IT distinction of data descriptions (state) and program code (process). By the term *process metadata*, we refer to the modelling and description of processes, regardless of whether these processes are implemented in program code or in instructions to statisticians carrying out the process by hand.

Metadata are used either to serve as *input for* statistical processes or to document *output from* these processes. In the former case, they *drive* the process, in the latter they *report on* the result of the process. We will use the terms *driving metadata* and *reporting metadata* to distinguish the two.

Reporting metadata are, like data, a product of the statistical process. They come into existence while or after a process runs. In contrast, *driving metadata* are needed for a process to run. Therefore, they must be supplied either in the design phase (explicitly predefined), or during the actual processing (on the fly).

Metadata-driven processes

Three expressions have entered the technical vocabulary of SN in the recent past: *metadata-aware tools, metadata-driven processes,* and *active metadata*. These terms turn out to refer to three aspects of the same concept.

The main idea is that the statistical process must remain under control of *driving metadata*. Metadata affect both the data that they describe and the processes that deal with the data. There are two different ways in which this control can be accomplished: either directly or indirectly. Indirect driving is achieved through instructions to statisticians on how to handle the process. Direct driving means that software can run the process without substantial human intervention: tools are *metadata-aware*. They take data and metadata as input. They produce output consisting of data and metadata. The resulting metadata contain a description of the output data, references to the processes that have produced the data, and information about the conditions in which these processes have run, including problems met in running the processes and possible deviations from the expected results. This makes it possible to generate *reporting metadata* (the documentation accompanying the data) automatically from within the process, instead of having to depend on humans to produce them in a separate process.

Metadata-aware tools are able to recognise metadata in data sets and access the data through the definitions in the metadata. This tight relationship between data and metadata is our guarantee that the metadata of the resulting data set are a faithful description of their data. In order to achieve this goal, the implementation of the metadata and the implementation of the tools have to be adapted to each other and support this type of interaction.

Figure 1: Metadata-aware tools access data and processes through metadata

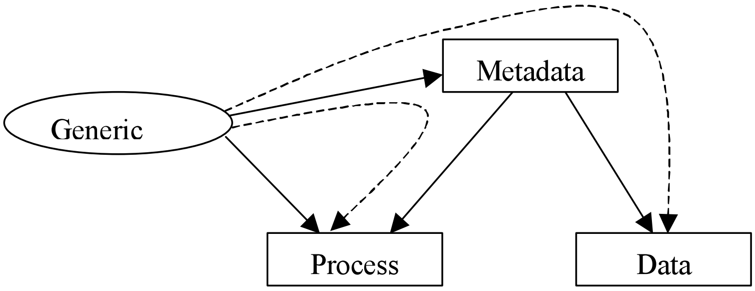
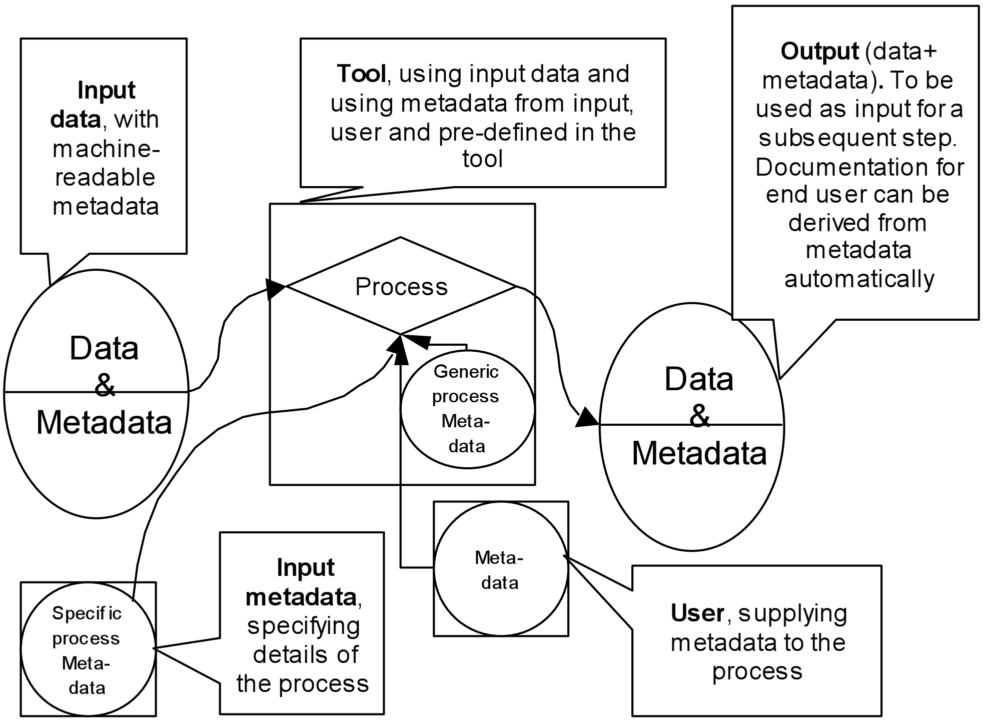


Figure 1 shows the relation between tools, metadata, data and processes. Straight-line arrows represent direct relations. Dotted curb arrows represent an indirect relation, which goes through metadata.

The relationship between tool and process in figure 1 deserves special attention. A tool has both a direct relation to a process and an indirect one. This is to express the fact that a process has both generic and specific properties. The generic properties are common to all instances of such a process. The specific properties differ from case to case. Let us look at the process of drawing a sample, for example. All sample draws have in common the fact that they take a population as input and produce a sample as output. These are the generic properties of the sample draw process. Sample draws, however, will differ from each other in the actual population extracted from, in the size of the sample, and in the sampling method. These are the specific properties.

Figure 2 illustrates the relation of a metadata-aware tool to different parts of the metadata. The tool is programmed to implement a process in a generic way and to accept different specifications for every instance of the process. Therefore, the metadata describing the process are split into *generic process metadata*, which are programmed into the tool, and *specific process metadata*, which have to be supplied every time the tool is used. The tool accesses input data through the metadata belonging to it, and

Figure 2: A metadata-aware tool


Input data, with machine-readable metadata
Tool, using input data and using metadata from input, user and pre-defined in the tool
Output (data+ metadata). To be used as input for a subsequent step. Documentation for end user can be derived from metadata automatically
Process
Data & Metadata
Generic process Meta-data
Data & Metadata
Meta-data
Input metadata, specifying details of the process
Specific process Meta-data
User, supplying metadata to the process

produces metadata to accompany the output data. Finally, if the tool implements a semiautomatic process, it must also be able to accept metadata supplied by the user.

A statistician should have the means to design a statistic exclusively in terms of statistical concepts and methods, and not in automation terms. In order to free the statistical designer from the burden of specifying *programs*, we need tools that implement methods in a generic way, and extract the actual tasks of a specific statistic from the information included in the metadata. Tools implementing *metadata-driven processes* in this way are beneficial for the quality of statistics for the following reasons:

- Tasks can be specified in statistical terms with no semantic bias from automation concepts.
- The metadata used in the design phase are the same as those used in the execution phase.
- Similar tasks are executed in a coherent, standardised way.
- The production process is reproducible.

Tools reflecting this concept are programmed to understand the metadata. They perform the tasks specified by the metadata, apply them to the data described in the metadata, and respect the constraints given in the metadata. They implement the generic aspects of statistical processes, and are completely under control of the metadata for the specific aspects. If this is the case, we talk of *active metadata*. Metadata are active if there is a systematic relationship between the content of the metadata and the behaviour of the tools, and if the behaviour triggered in different tools by a given metadata item is consistent.

The distinction between *active* metadata and *passive* metadata touches on the heart of the metadata strategy advocated here. The behaviour of the tools should be under control of the metadata. Metadata-driven processes are forced to behave as specified and cause the data to be as stated. Active metadata are technologically tied to the data they refer to, in such a way that any change in the metadata causes a corresponding change in the nature of the data, and data structures cannot be changed without a change in their metadata. In contrast, passive metadata can accidentally be changed without a corresponding change in the data.

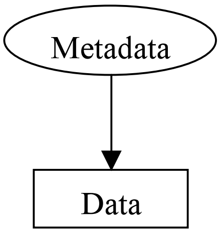
3. Goals for a metadata policy

Here is the most basic and brief expression of the goal our strategy seeks to meet: make sure that all data are, throughout their life cycle, completely, coherently and correctly described on a level of detail that satisfies the needs of users and supports the use of generic software. For metadata, this general statement can be elaborated by three specific goals, each of which represents a certain level of ambition:

1. **Faithful metadata.** The first of our three goals is to ensure that metadata give a correct picture of the data. In particular, the *reporting* metadata presented to the end users must correctly describe the data and the process that produced them.

This is illustrated in Figure 3, which shows data and metadata tied to one another, but free from any relationship to other metadata.

Figure 3: Faithful metadata provide a truthful picture of the data



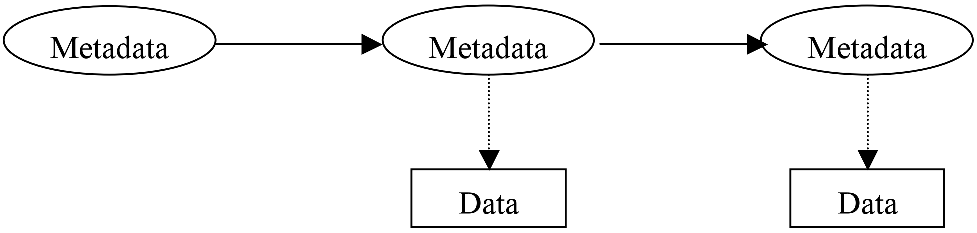
At this basic level of ambition, we do not pay attention to the *quality* of statistics. If the statistical concepts and methods that drive the process are not well considered and specified (for instance because they were made up on the fly), the resulting metadata will not be either. Still, if they give a fair description of these concepts and methods, however poor these may be, the primary goal has been achieved.

Of course, *we* are not satisfied, so an additional goal is introduced:

2. **Consistent metadata.** Our second goal is to ensure that the metadata *driving* the process and the *reporting* metadata presented to the end users are compatible with each other and correspond to well-considered concepts and methods, as defined by the process designer.

Figure 4 expands on figure 3 by adding a dimension to illustrate the relationship between the metadata of different phases of the process. The first occurrence of metadata in Figure 4 does not point to data.

Figure 4: Consistent metadata give the same picture of the data throughout the process



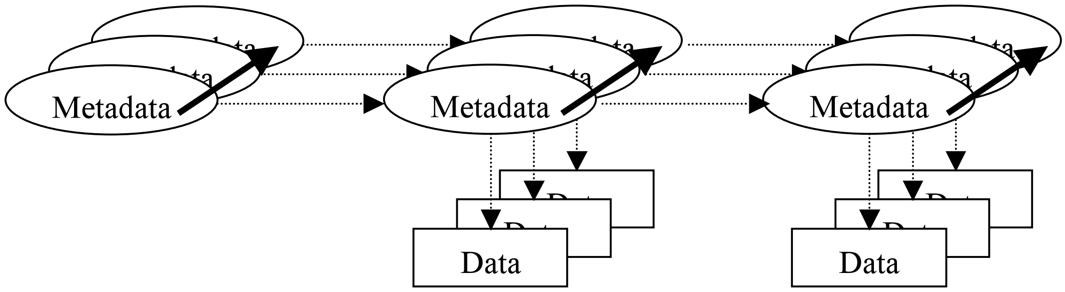
This illustrates the fact that Metadata appear before data (in the design phase).

Providing each individual statistical process with its own well-specified design is, however, still not enough: we want the data of *different* statistics to be co-ordinated and consistent. Therefore, we state a third goal:

3. **Standard metadata.** The third goal is to ensure that the metadata specified in the design comply with general standards.

To illustrate this, Figure 5 further elaborates on the previous figures by establishing a relationship between the metadata of different statistics.

Figure 5: Standard metadata give the same picture of the data of all statistics



It should be mentioned that there is an inherent contradiction between the traditional stovepipe model, in which every statistic is designed and carried out independently, and the co-ordination purpose of goal 3. Efforts to co-ordinate statistics without integrating them are costly and cannot be guaranteed to succeed.

4. A model for the statistical process

In section 5, we will discuss the implementation of the goals set above. Prior to this discussion, we need to present a general model for the processing of data and metadata.

The statistical process can be seen as a sequence of activities manipulating statistical data sets. In this context, a *statistical data set* is a unit of information consisting of *data* and associated *state metadata.* The state metadata describe the data in the data set.

In the process a statistical data set is produced, and later transformed into other data sets, until finally a data set is obtained that is suitable for disseminating the results. The transformation steps are called *tasks*. Each task controls the transformation from one data set into another. The transformation rules are laid down in the metadata driving the process. Tasks may also generate some of the process metadata. These metadata can take the form of parameters driving the next step of the process; generated metadata can also contain summaries of the task they were generated by, and can be used to monitor its quality. Examples of such quality indicators are the non-response rate or the number of detected errors.

Ideally, tasks are carried out by tools that ensure that the state metadata remain tied to the data. In practice, some tasks will be carried out by humans, and are therefore less easy to drive and document in a way that guarantees their faithfulness. For example, translation of concepts into survey questions will usually be a human activity that is hard to automate, and therefore remains hard to document properly.

Driving process metadata always contain at least three elements:

- Identification of the input data set;
- Identification of the output data set;
- Specification of transformation activities.

It is possible that a task uses more than one data set as input. For example, an adjustment-weighting tool uses a data set containing sample data, but also a data set containing the population data. It will also need to access metadata documenting the sampling frame and method.

Most tasks will also use information that does not belong to the input data sets, such as predefined metadata for their intended output. This means that in the transformation process, state metadata from the input data sets are combined with other metadata, and this leads to the state metadata of the output data set.

The metadata of each data set contain not only references to data descriptions, but also to the task by which the data set was generated. By following the references back to the start of the process, the history of the measurement of a variable can be retrieved and documented. This chain of relationships implements a virtual trail of steps that have been carried out. Audit trail documents can be extracted automatically from this information.

5. Implementing the three goals

Keeping in mind the blueprint of the statistical process described in the previous section, we will now discuss how this process should be arranged, in order to reach the goals specified in section 3.

The basic goal: faithful metadata

It is our conviction that the faithfulness of human-based metadata cannot be guaranteed. A shift to machine-based metadata is in our view a necessary condition for reaching the goal of faithful metadata.

Two key notions apply here:

- Use of metadata-aware tools and metadata in a machine-readable format.
- Automatic translation of machine-based metadata to human-readable metadata.

The concept of *metadata-aware tools* is important. The essence of the idea is that the statistical process is metadata-driven, being carried out by tools that place all their logic and functionality under control of *active metadata*.

Second goal: consistent metadata

The general idea is that the metadata used to control the tools should be derived as much as possible from a carefully developed design. Nothing should be left to chance; no metadata should have to be created on the fly, during the execution phase. The same machine-based metadata need to be accessed and operated on by the tools implementing all phases of the statistical process. This is particularly the case of the design phase, in which the designer specifies the metadata to be used in the rest of the process. Two recommendations apply here:

- Support for *metadata-driven processes.* Tools designed in such a way that their behaviour is controlled by the metadata.
- Make sure that the metadata needed in the consecutive stages of the statistical process "understand" each other, so that data and metadata can flow smoothly through the process.

The technical infrastructure on which the statistical process runs needs to support all the concepts relevant in the design of a statistical process: data structures, relations between data elements, constraints, statistical methods and the steps needed for their execution. The result of the design process must be available and understandable to the software managing the process. Active metadata are, in our view, the only way to achieve this goal. Metadata expressed in a machine-readable format, and tools that understand this format and execute the specified metadata can ensure that the processes being carried out faithfully reflect their design, and produce the data intended in the design.

Third goal: standardised metadata for both methods and concepts

Given the existence of a large number of statistics carried out independently (stove pipes), this goal has two dimensions:

- Within a particular statistic, the software tools supporting the different stages should be attuned to each other; they should "speak the same language". This is necessary in order to insure that the output of stage 1 (data and metadata) has the same form and meaning as the input for stage 2, and so on.
- Different statistics should make use of the same statistical concepts (e.g. a standard geographical classification; the standard variable *income*) and the same methods. Standardisation and co-ordination of statistics may affect their *content*, in the sense that concepts and definitions have to be adapted. The reward is that data resulting from different statistics are comparable and relatable. In the long run, there is also a positive effect on cost and efficiency.

The past has taught us that it is very difficult to really achieve standardisation and co-ordination in practice. First, it is obvious that both user needs and reality are too diverse to be fully caught in unique standards. Deviations are, therefore, to a certain extent unavoidable. Moreover – and this may even be a more important obstacle – moving towards standards means abandoning previously used statistical concepts and methods, thereby breaking time series. It is therefore not reasonable to expect short-term cost reductions from standardisation efforts.

The long-term objective is to offer the statistician a set of data processing tools that meet the following two goals:

- enforce the office's statistical program by limiting the choice to tools and metadata that comply with the standards;
- sell themselves by supporting statistical work in a way that is so attractive, efficient, and flexible that users will be willing to use them in spite of the reduced freedom associated with their use.

6. Implementation aspects

Metadata should be described as much as possible in a *formalised* way. This creates optimal conditions for the creation of metadata-aware tools. Here, also, there are positive side effects on the quality of statistical content. It turns out that this process provides us with a better understanding of what we are actually doing, and more insight in what we are doing wrong. Moreover, we have found that formalised metadata provide optimal support for presentation of statistical data, both on the input side (electronic questionnaires) and on the output side (electronic publications).

In order to be able to model the statistical process economically in terms of tasks and data sets, it is important to consider the following facts:

- the same task can run at different points in a process;

- the same task can be used in different statistics;
- the same data set can be used as input for different tasks;
- and different data sets can have the same structure.

To avoid repetition in all these cases, we do not want to describe tasks and data sets as they occur. We want to work within a framework that allows us to define *task types* and *data set types* in an abstract way. This will allow us to refer to those types whenever we need tasks and data sets that meet their definitions.

In such a system, a task definition is not a description of a specific task T that will (or has) run at a certain date D, takes some specific data set DS1 as input, and produces some other data set DS2 as output. A task definition describes a task type TT, which has a variable VD of type date, needs a data set of a certain type DST1 for its input, and produces a data set of type DST2 as output. DST1 and DST2 do not represent data sets, but their descriptions. VD is not a specific date, but a date property: it is an element of the description of the task, specifying that whenever the task runs, a date value has to be supplied. Moreover, TT is not the task itself, but a description of the task. TT, VD, DST1, and DST2 are all part of the metadata of a statistic (or of a collection of statistics).

A language for task definitions might describe TT as follows:

```
Define task type
    Name: TT
    Data:
        VD: Date
        Input: DST1
        Output: DST2
```

This only makes sense if the meaning of DST1 and DST2 are already known. Therefore, these descriptions also need to be available in the metadata repository.

Then, whenever an occurrence of task type TT is needed in the process, it is sufficient to define a task T by referring to TT:

```
Define task
    Name: T
    Type: TT
```

This means that the definition of task T is to be found in TT. This implicitly means that task T must know its date, and will work with two data sets, the descriptions of which are to be found in DST1 and DST2. The data sets need not be defined explicitly for task T: a reference to TT suffices.

We have so far described a task type TT, and one occurrence of this task type. At some point in the process model, we need to specify that this occurrence has to run, and indicate the actual data sets involved. This might be done with the following statement:

```
Run task
    Name: T
    Data:
        VD: D
        Input: DS1
        Output: DS2
```

This contains all the information needed by the system to take care of the following steps automatically:

- Make sure task T can run on date D. If D is in the past and T has not run yet, it is too late. If the definition of TT involves a date filter, D should fit this filter. Etc.
- Check that DS1 is of type DST1 – meaning that data set DS1 refers to data set type DST1 for its definition.
- Check that no existing data set is called DS2; then create a data set of type DST2 and call it DS2.

If the concept of active metadata is well implemented, these three steps need not be defined explicitly. They are a mere enforcement of the structures defined in task type TT, and can be inferred automatically. This is part of the generic behaviour of metadata-aware tools: check that the data and metadata are consistent with each other, enforce constraints defined in the metadata onto the data, and carry out the tasks defined.

Therefore, tasks need not define the variables involved explicitly. Instead, they refer to their own definition in the process metadata, which in turn refers to the state metadata of the data sets involved. Definition of variables takes place outside of the tasks using them. Inside the tasks, their definition is implicit (by reference).

This organisation of data and task definitions may seem too complex for the user to grasp easily. This, however, is not meant for human eyes. It is primarily meant for efficient access by software. One of the programs accessing such definitions must be a reporting tool designed to present this information in a format accessible to human readers. Deriving readable metadata in this way from the formal metadata driving the process is the unique guarantee that the two are consistent with each other.

An important spin-off of formally predefined metadata is a gain in efficiency both in the design phase and in the implementation phase. Reuse of previously defined task and data types accelerates design. Limiting the amount of decision-making that takes place after the statistical process has started running shortens the time needed to carry out the process and leads to more timely data delivery.

In practice, it is impossible to predefine all metadata in the design phase. First, some metadata can only be generated by the process itself, such as non-response rates. However, it certainly makes sense to decide beforehand what metadata will have to be computed during the process. In the execution phase, space will already have been reserved for this information, it will already have been defined, and the decision to compute this information will have been taken in advance. One can also predefine *targets,* like: “the non-response rate may not exceed x %”, or define actions to be taken if the computed value falls outside a certain range.

Second, some decisions can only be taken after certain information from the process is available. For instance, the choice of a weighting algorithm or the decision to carry out imputation can depend on the quality of the data collection process, which can only be determined after the collection has taken place. Such metadata emerge from the process as it runs. The moment at which such a decision has to be taken, however, as well as the rules governing it, can be determined during the design phase. The task of making this decision should be modelled in advance, so that all the required information can be available to the person in charge. Like predefined metadata, emerging metadata should be part of the model created by design.

7. People and machines

In the introduction we stated that metadata should evolve from their human orientation and become *machine-based* and *machine-oriented.* We have further stated that metadata should be modelled in an abstract way, should be stored in a machine-readable way, and should actively control the behaviour of metadata-aware tools. The reader may so far have gathered the impression that we have the ambition of fully automating the statistical process. This, however, is not what we have in mind.

There is, of course, much to win by keeping data and processes under control of one machine-based metadata system. The main profit is a better grip on the quality of the resulting data and metadata. There are, however, many points in the statistical process where human insight and competence cannot be replaced by software. The issue is how to integrate human expertise with a machine-based metadata system.

The solution to this problem lies in the separation of process description and process execution. The whole statistical process can be modelled in terms of subprocesses and of the data structures used by those subprocesses for their input and their output. This can be done regardless of whether a subprocess is carried out manually, interactively, or automatically. This global model, in turn, can be used for various purposes:

- It can produce documentation for data sets, for steps of the process, or for the whole process.
- It can build an automatic process from subprocesses for which automated implementation is available.
- It can monitor the execution of the process and produce reports.
- It can react to irregular situations by issuing warnings and reminders, and by blocking the execution of the process until errors are corrected or missing information is supplied.
- It can supply the information needed to produce an audit trail.
- It can trigger the execution of automatic steps by running the software independently.
- It can trigger the execution of semiautomatic steps and manual steps by sending e-mail messages to the people in charge of carrying them out.

In order to obtain the maximum benefits from such a system, it is necessary to concentrate decision-making as much as possible in the design phase. Statisticians know from experience which statistics require human decision-making after the process has started to run. They are able to specify such activities in advance and to describe them in terms of their place within the whole process, their relations with other parts of the process, the information required, and the conditions that have to be met by the results.

8. Metadata at Statistics Netherlands

We mentioned faithful, consistent and standard metadata as the primary goal of our metadata policy, and active metadata as the best way to achieve it. We wish to stress that the additional goals (a well-considered, neatly specified and formalised design) do not only have their own intrinsic value. They also create a favourable context for the successful implementation of metadata-aware software tools and contribute to the reliability of metadata.

In this section, we will present some of the metadata-oriented software of Statistics Netherlands (SN) in the light of the goals presented in this paper. This will allow us to take a critical look at what we have done until now, in order to unveil what is still required for reaching our goals.

The tools

The tools produced at SN were developed in a context in which co-ordinated, integrated and standardised metadata had not yet become an issue. The present challenge is to transform them progressively in such a way as to reach our goals on the long term without disrupting the work being done on the short term. This means creating new versions that maintain the present functionality and progressively introduce support for co-ordination and standardisation. In this section we will examine our two main tools, Blaise and StatLine, and the modules being developed for their modernisation, the Metadata Server (for storage) and the Cristal (for application logic).

Blaise

Blaise evolved from a system for defining questionnaires and producing interviewing programs in the mid-eighties to a metadata-driven statistical data processing system in the nineties. However, its roots are still visible in the fact that metadata for the input are better supported than metadata for the output. There is, for instance, no explicit support for aggregated data. There is no support, either, for modelling cells in a table. The textual information supported by Blaise is also typically directed to the input process: there are question texts, but no keywords or footnotes.

Blaise is a rule-based system. Data structures are accompanied by rules specifying the conditions under which data fields are relevant and enforcing relationships between the values of different variables (constraints and imputations).

Defining a statistical project in Blaise is done by specifying a data model containing all fields, constraints, automatic computations, and texts for questions and error messages. This information is then used by the different modules of the system to perform tasks like data entry, data editing, interviewing, and exporting metadata to other software, such as SAS, SPSS, or relational databases. The system is presently being extended to cover throughput processes such as macro editing, automatic imputation, and automatic documentation.

We consider Blaise as a successful example for the concept of active metadata. At the heart of a Blaise application is one global metadata structure (the data model), including constraints, imputations, texts and labels, which is accessed by all the modules of the system. This is the driving metadata. The data of a Blaise data set can only be accessed through their metadata. This ensures that the meaning of the data is the same for all parts of the system within the limits of one statistic. The conversion modules export both the data and the metadata in the format required by other software, ensuring that no human error has a chance of modifying the relation between data and metadata. The automatic documentation tool developed in the TADEQ project (Bethlehem, 1999) ensures that the reporting metadata faithfully reflect the driving metadata.

The Blaise system fully implements our first two levels of ambition. By encapsulating the data behind the specified metadata, it ensures that all users and all parts of the software get the same, consistent picture of the data. By requiring that the metadata be fully specified in the design phase, it ensures that the metadata involved in all phases of the process are consistent with the design.

However, by storing the metadata of separate surveys in separate files, Blaise does not enforce centralised and standardised metadata. Although Blaise supports modular development, reuse of building blocks, and links between surveys, it does not help designers of different surveys to co-ordinate their work. Therefore, Blaise falls short of our third level of ambition.

StatLine and StatBase

StatLine is the output system of SN, consisting of a hierarchically structured set of multidimensional tables, each of them describing a *theme* and together providing an overall picture of Dutch society and economy. It is now available both on CD-ROM and the Internet. Originally StatLine used to be updated directly from the source statistics.

In order to separate the structure of aggregates from their presentation, we developed the output database StatBase, designed to store all publishable aggregates, prior to their organisation in tables. This makes it possible to define StatLine tables without any limitations with respect to the boundaries of individual surveys. StatLine, however, will not be the only product to derive its data form StatBase: in the future no statistical data will leave SN, if they do not come from StatBase.

It is tempting to state that StatBase implements the concept of metadata-driven processes: statisticians supply the data and metadata of their publications and let the system do the rest. StatBase accesses the data through the metadata, and loads both into the database. StatLine uses the metadata in support of end-user activities: defining tables, searching and selecting.

The system, however, has no control on the *quality* of the metadata submitted to StatBase. We do not have standard aggregation and integration tools yet, so our statisticians are free to use the software of their choice for the production of data sets meant for publication. Therefore, the StatLine database contains no information about the origin of the data, in terms of input data sets and transformations having led to the data sets from which the publications are derived. It can be said that StatBase endeavours to implement our third ambition level without devoting any attention to the fist two levels. The quality of the metadata available in StatBase and used in StatLine depends on the quality of the work done in the preceding phases, but StatBase has no access to this information.

The databases

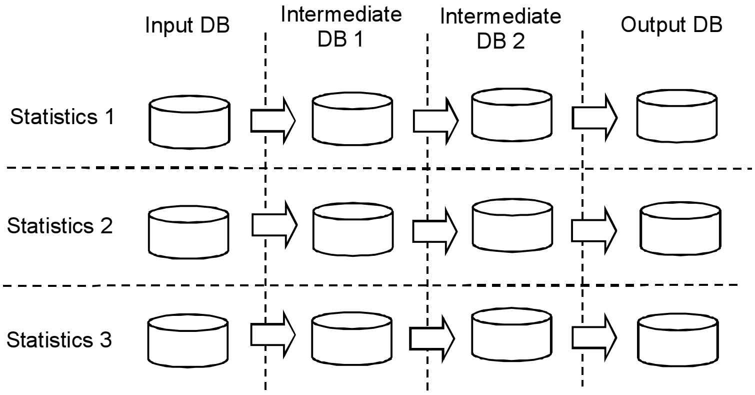
In order to migrate smoothly from the old survey-oriented "stovepipe" model of the statistical process to a more process-oriented model, the numerous data storage systems previously used by different statistics were replaced by a limited number of generic databases. The word database must be taken here in a broad sense. There are a number of storage formats, ranging from text-based data files and proprietary formats to relational databases and data warehouses. It is necessary to take various formats into account in a context in which data are received from various sources, and in which various programs are in use.

The aim is to co-ordinate statistics from the point of view of the statistical process, and to use this co-ordination in order to progressively co-ordinate and integrate the data themselves.

The four-database model

In order to allow for consistent statistics in the output database, the input and the intermediate databases have to be integrated across the borders of the stovepipes.

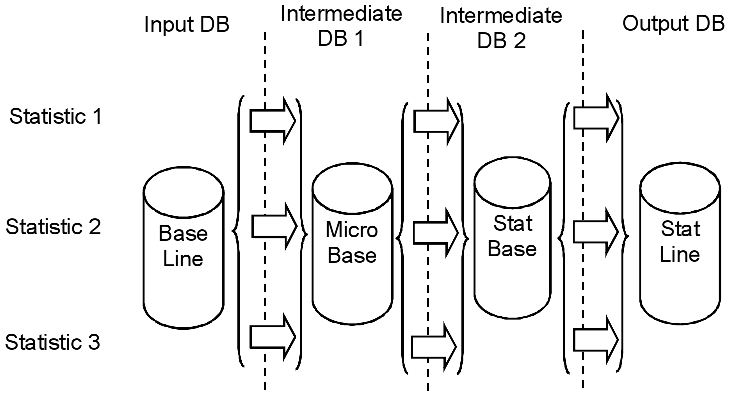
Figure 6: Schematic picture of three independent stovepipes



At SN this integration is called *vertical integration.* It has resulted in a process supported by four global databases:

- **BASELINE** joins data from all input databases. It holds all data as supplied by all data sources such as administrative registers or responses from direct surveys.
- **MICROBASE** joins data from all first intermediate databases. It contains the data as they result from imputation, translation and micro-integration.
- **STATBASE** joins data from all second intermediate databases. It contains aggregated data at the level of populations and sub-populations after estimation and macro-integration. STATBASE holds all *publishable* data produced by SN.
- **STATLINE** joins data from all output databases. It shows the total output of Statistics Netherlands as a structured set of multi-dimensional tables, each covering an area of societal interest. STATLINE is published on CD-ROM and via the Internet.

Figure 7: A schematic picture after vertical integration with four databases



The data modelling of a vertically integrated database is done in several stages:

- the *vertical data collection* stage, in which all data structures from different stovepipes are put together in one big but highly fragmented database with a simplified data model;
- the *vertical data confrontation* stage, in which the defaults of the previous stage are discussed: information structure which is too fragmented, lack of consistency;
- the *vertical data restructuring* stage, in which data model enhancements result in a defragmentation of information and consistency improvements.

The last two stages are repeated in a loop until a satisfactory result is reached.

Common metadata

The previous section reveals a major weakness in the statistical systems originally developed at SN: input and output data and metadata are treated by systems that do not have a common metadata approach. So there is a point between the two where data and metadata undergo a change that escapes control from both systems. It is therefore impossible to guarantee that data and metadata remain synchronised (ambition level 1). It is not possible either to ensure that design decisions are implemented in a consistent way under both systems (ambition level 2). If it is presented with metadata that are standard (ambition level 3), StatBase will maintain this quality, but it cannot influence its input.

Therefore, although our aims were partially met by the two systems, our statistical process as a whole needed to be revised. There was an urgent need for a system to record and process metadata in a uniform way for all our tools. We needed to be able to define concepts and variables clearly and unambiguously within the framework of the Sundgren data model. We also needed to ensure that these definitions were understood uniformly by all our software, independently of specific statistics and of processing phase, and to take advantage of the good support offered by development tools since the late nineties for separation of logic and storage aspects. This lead us to develop a *Metadata Server* for the storage and a logic model for a run-time component called *Cristal.*

The Cristal model

The ideal situation of Figure 7 is only half of the total integration process. At that stage of integration, the four integrated data models are constructed almost independently. But of course the models for the data in each database are not independent.

Similar information is used in all four databases. Each of the four databases gives a specific structure to the subjects in the world as perceived by SN and does it on a different level of detail and from a different viewpoint. But to be mutually consistent, they should picture the same world in the same way. Therefore a unique data structure, independent of the process stages at SN, is needed for the description of the common world.

Integrating data structures taken from different stages in the production process will be referred to as *horizontal integration.*

For the horizontal integration modelling the same stages can be distinguished as for the vertical integration modelling: the *horizontal data collection*, the *horizontal data confrontation* and the *horizontal data restructuring* stages.

Intersection of Concepts and the Cristal Model

During confrontation and data restructuring stages, the concepts used in the different databases are examined. Some concepts are specific for the stage in the production process, others are generic. The overlapping concepts will be called ‘intersection concepts’ here; indeed these concepts are most subject to confrontation and restructuring. Figure 8 represents the concepts used in the four databases with circles and emphasises their intersection.

The following bullets describe some *specific* concepts for the stages in the production process. These concepts lie outside the intersection shown in Figure 8:

- **BASELINE** (input information):

  Register, Question, Questionnaire, Questionnaire routing, Interviewer, Interviewer instructions.

- **MICROBASE** (intermediate micro information):

  Outlier, Imputation, Imputation Rule, Micro Edit, Macro Edit, Weight.

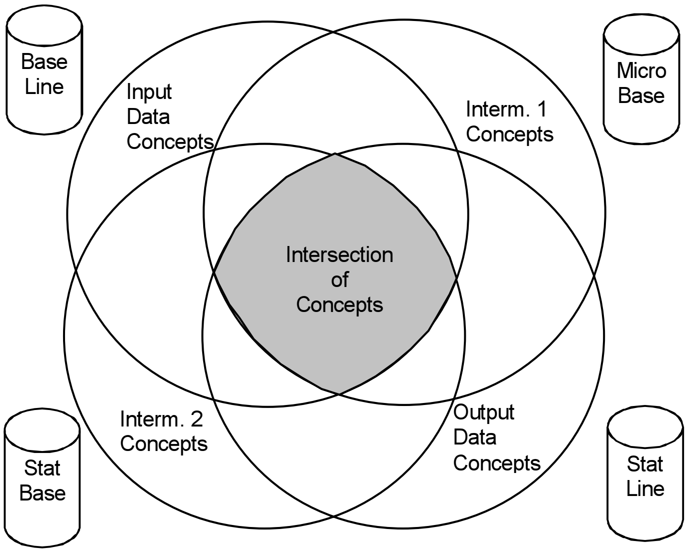
- **STATBASE** (intermediate macro information):

  Aggregation, Estimation, Weight

- **STATLINE** (output information):

  Cubes, α-, β-, γ- and τ-variables, Variable groupings, Table, Table layout, Grids, Rows, Columns.

Figure 8: Intersection of concepts from four databases



The following *generic* concepts are used in all four databases and lie in the intersection shown in Figure 8:

Object Type, Object, Population, Variable, Classification, Property and *Value.*

The Cristal model is a data structure for the generic concepts (in the intersection). It is meant to tie loose ends together, both in vertical and horizontal integrations. The Cristal is not visible in storage or presentation structures. It is a purely logical model, implemented in a run-time component.

The Metadata Server

This is a set of centralised metadata repositories containing all the items our data sets refer to. The first module of the Metadata Server, the Classification Server, has been implemented, and is successfully being used by a range of applications. Ideally, this should ensure a uniform approach to both flat code lists and hierarchical classifications throughout the statistical process. In fact, we found out that the availability of a central repository is only the first step. All relevant applications need to access it before it can be used to its full value. This is the phase in which SN is presently engaged. New applications access the Classification Server. Old applications are being redesigned to use it. This is particularly the case of StatBase, the tool from which the StatLine database gets its data and metadata. All statistics have to use the Classification Server before they can be loaded in StatBase. This is true whether they use it or not in previous phases.

When we reach a situation in which all applications refer to the Metadata Server for their metadata, we will have implemented the whole range of our goals at the storage level.

References

Bethlehem, J.G. & W.J. Keller (1990), *The Blaise System for Integrated Survey Processing*, Statistics Netherlands.

Bethlehem, J.G., F.M. Kellenbach & W.J. Keller (1991), *Computer Assisted Statistical Information Processing at the Netherlands Central Bureau of Statistics*, Voorburg, Netherlands Central Bureau of Statistics, 1991.

Bethlehem, J.G. (1999), "The Routing Structure of Questionnaires", *Proceedings of the Third ASC International Conference*, p. 405-414.

Blaise III Developer's Guide, Statistical Informatics Unit, Statistics Netherlands, 1994 (Blaise III 1.0) and 1996 (Blaise III 1.1).

Bracht, Eric van, "CRISTAL, a Model for the Description of Statistics", paper to be presented at NTTS/ETK 2001.

Gillman, D.W. *et al.* (1996), "Design Principles for a Unified Statistical Data/Metadata System", *Eighth International Conference on Scientific and Statistical Database Management*, pp. 150-155.

Karge, R. (1999), *ComeIn – Common Metadata Interface*, Berlin, Run Software Werkstadt.

Kent, J.-P. & G. Razoux Schultz (1993), "Recent developments in Blaise", in P. Debets *et al.*, *De computer als veldwerker: verzameling, invoer en kwaliteit van gegevens*, Amsterdam, Informatiseringcentrum, pp. 91-110.

Kent, J.-P. & M. Schuerhoff (1997), "Some thoughts about a Metadata Management System", *Ninth International Conference on Scientific and Statistical Database Management*, pp. 174-185.

Olensky, J. (1996), "Practical Problems of Implementing Metadata Standards in Official Statistics", *Eighth International Conference on Scientific and Statistical Database Management*, pp. 130-147.

Richter, W. (1996), "The ABS Information Warehouse – Present & Future", *Conference on Output Databases*, Voorburg, November 1996, pp. 11-18.

Sundgren, B. (1973), *Infological approach to databases*, Urval Skriftserie 7/1973, Stockholm, National Bureau of Statistics.

Sundgren, B. (1977), "Meta-Information in Statistical Agencies", *Conference of European Statisticians*, ISIS'77 Seminar.

Sundgren, B. (1981), "Statistical Data Processing Systems – Architectures and Design methodologies". Invited Paper for the Golden Jubilee Celebration Conference on "Statistics: Applications and New Directions", Indian Statistical Institute, Calcutta, December 1981.

Sundgren, B. (1991a), *Towards a Unified Data and Metadata System at the Australian Bureau of Statistics* – Final Report.

Sundgren, B. (1991b), *Statistical Metainformation and Metainformation Systems*, R&D Report Statistics Sweden.

Sundgren, B. (1992), *Organizing the Metainformation Systems of a Statistical Office*, R&D Report Statistics Sweden.

Sundgren, B. (1993), *Guidelines on the Design and Implementation of Statistical Metainformation Systems*, R&D Report Statistics Sweden.

Willeboordse, A. & W. Ypma (1996), "From Rules to Tools – new opportunities to establish coherence among statistics", *Conference on Output Databases*, Voorburg, November 1996, pp. 50-57.

Willeboordse, A. & W. Ypma (1997), *Meta Tools in support of a Corporate Dissemination strategy*, Statistics Netherlands, Research Paper No 9839.

About the Author

Jean-Pierre Kent works in the Methodology and Informatics Department of Statistics Netherlands.

THE CHALLENGE OF THE INTERNET
Edited by A. Westlake et al
Compilation © 2001 Association for Survey Computing

An Open Data Model for the MR Industry: The Challenge and the Promise

Peter Andrews

Abstract

Historically, the survey research industry and its computer software suppliers have been forced to accept barriers to the interchange of survey data between software systems and operating environments. These barriers, arising in part from disparate formats for handling the unique requirements of survey data, have been handled primarily through import and export programs, usually of limited effectiveness, and substantial production delays and costs are often incurred as a result. In addition, a core set of common functionality is often created in parallel by each software supplier or division, which affects software costs and ability to adapt quickly to change.

The fundamental need is not only for a medium of interchange for study information and respondent data, but for a means of simplifying the creation and linking together of application programs across a diverse range of functionality. This paper will provide an overview of an effort by SPSS MR to create an open data model that provides an interchange medium, a published API (Application Programming Interface), and extensive functionality to simplify the development and integration of applications for survey research.

Keywords

Standards, open, survey, market research

1. Introduction

Historically, the survey research industry and its computer software suppliers have been forced to accept barriers to the interchange of survey data between software systems and operating environments. These barriers, arising in part from disparate formats for handling the unique requirements of survey data, have been handled primarily through import and export programs, usually of limited effectiveness, and substantial production delays and costs are often incurred as a result. In addition, a core set of common functionality is often created in parallel by each software supplier or division, which affects software costs and ability to adapt quickly to change.

While significant headway has been made against the basic interchange problem by standards groups such as the Triple-S Committee, the challenge is in the meantime evolving. The survey research industry is beginning to see growing client demand for ever more complex analysis and data mining, for customer relationship management, for web-based publishing of interactive databases, and for executive information systems and marketing information portals that integrate multiple data streams. As we rise to meet these needs, we encounter fundamental inefficiencies arising from the difficulty of integrating applications from multiple vendors to create comprehensive solutions for clients.

The fundamental need is not only for a medium of interchange, for study profile and respondent data, but for a means of simplifying the creation and linking together of application programs across a diverse range of functionality.

The challenges include both logistical and industry issues: on the one hand, the need to handle the requirements of survey data, as well as practical issues that include hierarchical data structures and the tracking of changes over time in longitudinal survey instruments; and, on the other, to achieve consensus among industry players and software providers on the benefits to all of adopting a "lingua franca" that will allow applications from competing suppliers to share information simply and openly.

This paper will provide an overview of an effort by SPSS MR to create an open data model for use by the survey research industry. It will review some of the challenges; the potential benefits for agencies and their clients, systems suppliers, and the industry as a whole; and learnings gathered to date from this extensive undertaking.

Design Goals and Benefits

With these challenges at hand, and with a perceived industry need for a shared, open solution to alleviate the frustrations described above -- and to facilitate the creation of larger, more complex next-generation applications -- the need seemed clear for a technical solution that would meet at least the following design goals:

- To develop MR applications independent of data storage formats
- To make it easy…
  - to read & write MR data and metadata
  - to develop new MR applications faster
  - to integrate SPSS products end-to-end
  - to leverage or integrate 3rd party tools

To do this, the Data Model must be than just a repository for survey data. It must provide functionality to allow applications to handle unique aspects of MR data, including multi-response questions, hierarchical datasets, and tracking of study versions, simply and quickly, so that these wheels need not be reinvented for each study. And it must provide an API that simplifies application programming.

The key benefit for the survey research community would lie beyond simple interchange of data, in customization of systems, products and business processes:

- The opportunity to re-engineer and broaden the business
- The control to customize the research process
- The independence to differentiate through innovation
- The power (speed, quality and efficiency) to reap quantum productivity gains

An early, simple real-world example might help to illustrate this last point. The construction of paper questionnaires for scanning often involves some fairly extensive formatting, as well as a time-consuming step in which fields on the paper are mapped into the scanning software. We've combined third-party software from multiple vendors with our data model and a proprietary application to create a solution that automates much of the formatting – and reduces the mapping process from one that often requires a day or more, depending on the questionnaire, to a matter of seconds. Further, changes to the questionnaire are automatically handled, both in the formatting and in the mapping to the scanning software. This saves multiple iterations of the manual mapping process, and allows agencies to be far more flexible and responsive.

A research agency, using these standard tools, could do the same thing.

2. Challenges

In addressing this ambitious agenda, the design team faced some key challenges in meeting its design goals while taking into consideration the following:

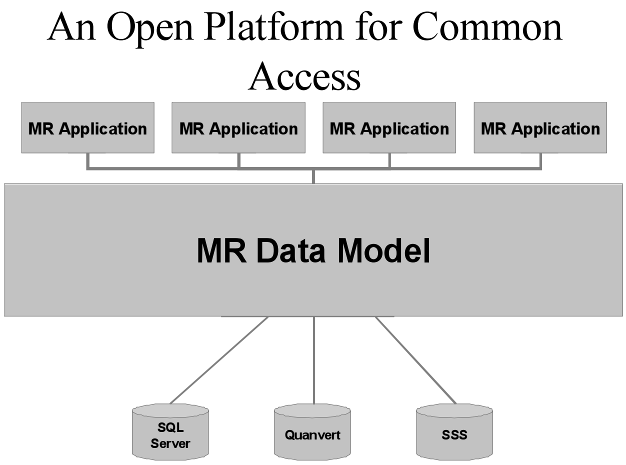
- Implementing unique MR case data structures while taking advantage of standard database environments and applications
- Fully allowing in metadata for the complexities introduced by factors such as tracking and versioning, global research, and hierarchical datasets
- Accommodating the widest possible range of data implementations, including legacy storage formats
- Designing simple-to-use API's that would facilitate complex functionality
- Choosing industry-standard technologies that would
  - Facilitate the use of the widest range of new technologies, by us and others
  - Allow the use of best–of-breed software "building blocks", so that we and others develop only what's unique to our industry or application
  - Provide a compatible solution for the widest number of user organizations and industry platforms
  - Leave us prepared for future technology change
- Making the benefits of industry adoption compelling, and the barriers minimal
- Overcoming tactical thinking, vs. strategic, both inside and outside SPSS -- e.g., proprietary & closed vs. open, industry standard & published.

3. Meeting the Challenge: The Data Model

Fundamentally, the data model is a set of data/metadata API's to enable applications to:

- Author
- Collect
- Analyze
- Display
- Print
- Import/export

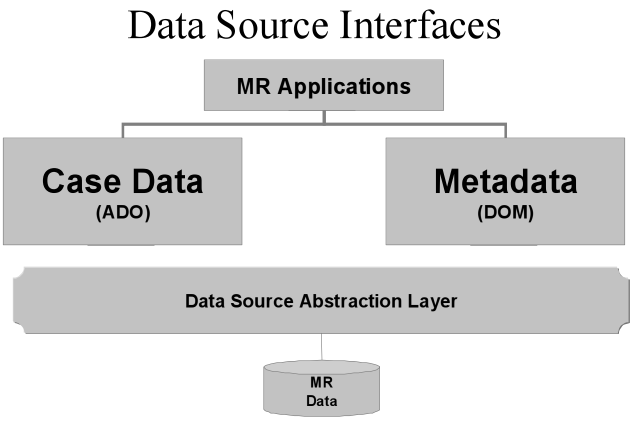
…in a survey research environment.



The data model hides the implementation details of underlying, package-specific storage architectures, so that programmers of application programs need not even be aware of them. Each user program is written using a standard interface. Thus, the application can be insulated from any changes to the underlying data storage format, software environment, and even hardware platform. Further, the data model provides functions to handle common tasks such as simple aggregation and the comparison and manipulation of multi-reponse data, freeing the programmer to focus on the unique aspects of the application.

The data model API's are based on Microsoft technology, and make it straightforward to create custom applications in the Windows/ASP environment.

Case Data



Case (respondent) data is read and written using the standard ADO interface, using a subset of standard SQL commands, extended with a set of functions which provide such features as the ability to access and manipulate multi-response ("multipunch") category variables. Lower-level OleDB and ODBC interfaces are also planned.

Regardless of the underlying data storage format, the ADO provider exposes case data to the consumer application in a very simple virtual relational table. (Additional tables are used when representing hierarchical, or 'levels' data.) The columns presented in this 'VDATA' table correspond to the variables (e.g. questions) requested in the application's SQL query, and the rows correspond to respondent records. The results of a query are presented by the provider as a rowset of the selected data. For example, the following query would produce the table shown below:

```
SELECT tested, color, advert, advert_other, product from VDATA
```

tested	colour	advert	advert_other	product
2	/3/	1		2
3	/2/	1		2
3	/1/2/	1		1
1	/2/	1		2
2	/2/	1		2
3	/3/1/	2		10
1	/4/	7	Poster	10

In this example:

- Each row contains response data for a single respondent
- The **tested** and **product** columns contain the data from single-response categorical variables, which the provider presents as integer values.
- The **colour** column contains the data from a multiple response categorical variable, indicated by the slash-separated values. These values can be presented in order of mention by the respondent.
- The **advert** and **advert_other** columns contain the data from a single response categorical variable, 'advert'. The last row includes an associated open-ended response, and the text for this "other mention" is recorded in the **advert_other** column.

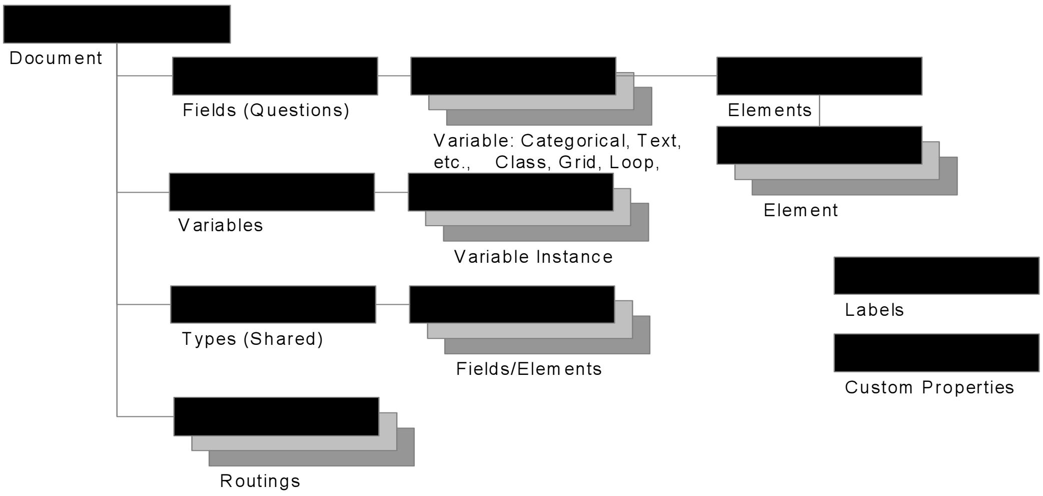
Metadata

The comprehensive metadata object model is designed to accommodate all required project information other than the case data – that is, information about the study, rather than the respondents

- The text of questions that were asked and the names of variables used in analysis.
- Other source material used in a survey, such as multimedia presentations.
- Versioning data that allows the reconstruction of all historical versions of a project.
- Alternative texts for use in different contexts (e.g. data collection and analytic environments).
- Translations of texts into any number of languages.
- The structure of the data, such as loops, grids, routing and derived variables
- Hierarchical data structures
- Any other details about the case data that you may want to record.

This information is stored in an XML file. The data model interacts with the XML document through another Microsoft standard, the Document Object Model (DOM), and provides a straightforward API for application programs.

Metadata Object Model (subset)



In the chart above…

- "**Fields**" can be thought of as containers for survey questions or groups of questions (e.g. grids, rotation groups)
- Each "**Variable**" contains the detailed properties for a question or other variable
- "**Elements**" are members of response lists
- "**Types**" are shared response lists, which may be used by multiple questions.
- "**Labels**" hold any other text, and are stored in a multidimensional array according to multiple attributes: user context (e.g. question prompts, interviewer instructions), language, data collection mode, and application context (e.g. data collection vs. analytic database).
- "**Custom Properties**" are user-defined fields used to extend the metadata content. These can be used, for instance, to hold column locations, tabulation instruction tags – or whatever unique information your authoring tool, questionnaire database or other application requires.
- "**Routings**" currently hold question order for each version (e.g. CATI, Web, WAP) and simple questionnaire logic such as skips and filters. Full, "CATI-style" questionnaire logic is planned for future implementation.

Information on questionnaire versions is also maintained by the metadata model. Once a version number is specified by the end-user application, the current view shown to the application will be of that version alone. This permits applications that must be aware of changes to questionnaires over time, such as tabulation databases and OLAP environments, to access exactly what was asked in each wave, for instance, of a tracking study. A synchronization process, initiated as each version is finalized and "locked in", allows each version to store new, unique information separately, while incorporating information common to multiple versions by reference.

How *Easy to Use* Is It, Really?

The box below shows the six lines of Visual Basic code required to open a metadata object, query it and print out the names of the question fields. Compare this with, say, the time required to write a program to extract the same information from a Triple-S metadata file – or from a closed architecture such as Quanvert, where it can't be done at all

Accessing Metadata

```
' VB snippet to print variable names from an MDM Document

Dim Document As New MDMLib.Document
Dim V As MDMLib.VariableInstance

  Document.Load "MyMetadata.mdd"
  For Each V In Document.Variables
    Print V.Name   ' print names of VariableInstances
  Next
```

Handling Diverse Data Formats

The "abstraction" layer, which makes diverse data and metadata storage formats accessible through the data model API's in real time, at the individual transaction-level, is achieved through the creation of "data source components", or DSC's. These are programs (COM objects), individually written for each storage format, which serve as adapters or translators. Where the high level of integration provided by data source components may not be required – where batch conversions are sufficient, for instance – simpler programs using the API can be written to import studies from proprietary formats into the Dimensions Data Model, and thus to the many data formats supported by the Data Model.

Within the SPSS product line, DSC's have been or will be made available for Quanvert data, SPSS .SAV files, mrInterview's relational database, and so on. A software developer's library is being made available to assist other software suppliers, research agencies with proprietary databases, and others in creating these components and with using the API. Once a DSC is complete, the agency or supplier will have the opportunity, over time, to integrate proprietary applications at the transaction level with the full line of SPSS products and with those of other suppliers who have written DSC's – and import and export programs for these packages should be a thing of the past.

4. Open Architecture

The data model is "open" on two levels: first, it is built around open, industry-standard technologies; and second, it provides an open, published interface (API) for the use of one and all. The first allows DM-enabled programs to interact smoothly with a large body of other standards-based software, allowing them to be combined easily as components of more complex systems. The second allows others outside SPSS – clients, partner companies and even competitors – to easily take advantage of the benefits offered by the data model.

Component Technology

The data model is built around industry-standard, open software technologies. From a practical standpoint, the data model's functionality is provided by a set of COM components, implemented as dynamically linked libraries (DLL's). COM is Microsoft's Component Object Model, which allows programs to expose a program interface to other programs. This enables Windows applications to open and control sessions in other programs such as Excel, Word and PowerPoint. Many survey research agencies use this functionality in MS Office to automate tasks such as report and graphics production. The COM standard facilitates the combining of "component" programs such as the data model, the Office products and many others to build applications.

Your proprietary applications can control the entire process in a straightforward way, using Visual Basic or any of the languages that work with these standards. Thus, our products and others can be combined and extended to create your own proprietary applications – you can add your own unique value, and your own look and feel, combining and adding onto these building blocks. You need not go to the expense and trouble of recreating core functionality – your investment can be focused on those specific areas where your products break new ground.

The use of open standards provides a tremendous efficiency benefit at every level – whether you're simply loading a dataset into Excel with a few lines of VBA, or you're creating a complex tracking environment or a marketing information portal. Moreover, because you can work with survey data independently of the underlying storage format, you are not tied to any one product or software supplier. The result is greater flexibility and improved business processes.

Publication

In order to encourage adoption of the data model, the API's are being published and a developer's kit for the data model is being made available to the industry at no charge. The data of this conference happens to coincide exactly with the public release of the Dimensions Development Library. Only if the data model – or something like it -- is adopted broadly, however, will it yield all of its potential benefits.

To this end, SPSS is supporting industry standards efforts wherever possible. In the MR area, our director of development has been working with the Triple-S committee, and will contribute in any practical way to the converging of our data model with this published standard.

On another front, the Object Management Group (OMG) has published the Common Warehouse Metamodel (CWM), a specification for metadata interchange among data warehousing, business intelligence, knowledge management and portal technologies.

This is a comprehensive standard. It includes relational database structures, OLAP storage, xml structures, data transformation specifications, application of analytic techniques, specification of presentation formats for information, and operational parameters and metrics for data warehouse operation.[13] SPSS has business intelligence products in areas such as analytical CRM, business metrics, and data mining, which currently utilize proprietary, xml-based metadata formats. SPSS is currently working on implementation of a metadata repository that supports CWM-based XMI interchange, which would bridge existing internal standards. While much existing metadata maps fairly easily to the CWM standard, there are places where the standard's definition of variables would have to be extended to support required functionality in the SPSS products. The tentative plan is to adopt this CWM-based repository when it becomes available. If this does in fact happen, then the MR division intends either to adopt the same metadata model or to link with it transparently.

5. Conclusion: Strategic Thinking

Our goal at SPSS is to build on the deep, long-term partnerships we have with our clients. We must earn and sustain their trust -- they must know that we as a company are out to help them deliver the most effective possible solutions for their customers. SPSS has a clear-cut commitment to open standards and systems, so as to allow clients maximum flexibility to extend and combine our products to create differentiated solutions that add value in turn for their customers. More and more, SPSS products are offered both as turnkey solutions and as components, developers' tools that support our clients in this endeavor.

This isn't the traditional thinking in this industry. Companies have felt that protecting their secrets meant protecting their client bases. The problems that result have become painfully clear to SPSS as, following our acquisition of three MR software companies, we've had to rationalize, integrate and maintain within our own product line three parallel lines of traditional MR software that didn't really talk with one another, and somehow to integrate them with our wide array of analytic, business intelligence, graphics, and other applications. One of these companies, Quantime, was well known for closed, proprietary systems such as Quanvert.

Today, though, it is important to leave this kind of tactical thinking behind. As the MR industry changes, as it faces threats from consultancies and from the do-it-yourselfers, as demand emerges for more and more sophisticated data warehousing, analytic and marketing information applications, the stakes are far too high. Our clients will be – and should be -- intolerant of company policy that "traps" them within a product line.

It is important to all of us, as an industry, to think strategically about open systems and to support industry standards. By adopting standard technologies, by publishing our API's, and by working with industry standards organizations, we can as an industry help our clients to compete effectively in this changing marketplace. And when our clients win, we win.

[13] The CWM was developed by a group of companies that include industry leaders such as IBM, Oracle and Hyperion. This standard may be merged with another standard from the Meta Data Group, until recently a separate coalition that included Informatica, Microsoft, and SAS. If and when the work of merging these standards is complete, the resulting specification will be issued by the OMG as the next version of the CWM. A single standard should allow users to exchange metadata between different products from different vendors with relative freedom.

About the Author

Peter Andrews is Director of Business Strategy in the Next Dimensions Team at SPSS MR. A key component of his role is to promote the wide acceptance and use of the SPSS MR Data model by customers, future customers and partners.

Peter has worked in the MR industry for 20 years. He joined SPSS from Audits & Surveys Worldwide (part of the United Information Group), where he was Senior Vice President in charge of New Product Development and, previously, worldwide CIO.