Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

WWW Surveys

Zenel Batagelj and Vasja Vehovar [1]

Abstract

The age of the Internet provides a new and powerful tool for modern survey data collection: the WWW survey.

The major advantages of WWW surveys are the speed and extremely low costs of data collection. For these reasons a rapid development of the WWW survey methodology and the growth in use of WWW surveys can be expected. Now, the WWW surveys are mostly used for surveys of populations with high Internet coverage (computer users, institutions, companies, organizational research, and certain professions).

In this paper technical and methodological aspects of WWW surveys are discussed. Technically the development and current stages of WWW surveys are presented. From methodological point of view, different aspects like the sampling problems, the promotional activities, the use of incentives, the duration of the survey, the length of the questionnaire, and the form of the survey are discussed.

The discussed aspects of the WWW surveys are illustrated using the data of the national WWW survey in Slovenia. The survey was conducted in 1996 (n=1200) and 1997 (n=3500).

1 Introduction

Interviewing over the Internet has developed in an extremely short period of time and is still in a phase of very intensive development. Therefore articles in this field of research are still scarse. We could say the same for presentations at professional conferences – although attractive and quite well attended – they are relatively uncomplete and moderate in contents. At the end of the year 1997 program modules for WWW interviewing of software packages were in beta version phase at the most, now they are included as regular modules in software packages.

[1] Faculty of Social Sciences, University of Ljubljana, P.O. Box 2547, 1001 Ljubljana, Slovenia; e-mail: zenel.batagelj@uni-lj.si and vasja.vehovar@uni-lj.si.

The authors would like to thank Willem E. Saris and Irene Hanson Frieze for their comments on drafts of this paper.

2 Internet as a surveying tool

As mentioned before, the development of interviewing over the Internet was very rapid – as is the case with the Internet itself over the last 4 years. Nevertheless we can establish that the present stage of development of the Internet interviewing has reached a technical level which includes most of the knowledge from the field of Computer Assisted Data Collection (CADAC), such as complex branching, validation of responses, randomization of questions and answer categories (for details see Saris, 1991). Kottler (1997) presented some reflections on the development of Internet based data collection. In this section some further extensions of Kotler's classification are presented.

In general, Internet based interviewing can be divided into two groups: e-mail and WWW based interviewing. From technical point of view and also from respondent's point of view, WWW based interviewing can be further extended into HTML based interviewing and JAVA applets and Javascript based WWW interviewing.

2.1 E-mail based interviewing

Interviewing by e-mail represents the oldest and the most simple form of interviewing over the Internet. E-mail interviewing was used long before the Internet was introduced, and so has e-mail based interviewing.

2.1.1 E-mail-only based interviewing

In this mode of e-mail interviewing, respondents receive e-mail in a form of a text into which they insert their answers. A weakness is mainly in non-standardized program equipment of respondents and in a complicated procedure of data input from the respodent's point of view. Transfer of answers into a specific datafile can also be very problematic. Programs which arrange the answers into databases do exist, but because of errors at the input stage, they fail frequently. This means that some manual extra work must be done.

2.1.2 Disk by e-mail

The idea of the »disk by e-mail« is simple: a computer program which is basically a questionnaire (computer self-administered interviewing) is attached to e-mail and saved on the respondent's computer. The respondent runs the program, answers the questions, and e-mails the file with the answers. Such a procedure is complicated compared to the previous one, but has great advantages because computer assisted

self-interviewing which enables complex skip patterns, mixing of questions, questions control etc. can be used.

2.2 HTML forms-based interviewing

With the rapid diffusion of WWW use among Internet users and the HTML standard which enables simple and standarized input of answers, interviewing was quickly transferred to WWW.

From a technical point of view there are three significant issues in the HTML forms-based interviewing.

Questionnaires as static or active WWW pages

When a web page that includes a questionnaire is called by the browser, the questionnaire can be a static or an active WWW page. The later are pages that are the same for each respondent. The former are those that are actually generated by a computer program. In such cases for some applications, very important data can be collected :

1. the name (DNS or IP) of the computer used when filling the questionnaire,
2. the browser used,
3. the operation system used,
4. the Web adress, from which respondents came from,
5. starting time of the interview,
6. in special cases, also respondent's e-mail address (ie. text browser Lynx).

Browser activities in data-entry period

With the introduction of Javascript, WWW pages became active without interaction with server. In a context of questionnaires, this means that with a help of Javascript consistency checking can be done without interaction with WWW server.

Posting the data

When the questionnaire is filled out, a browser has to send the data. This can be done in two different ways:

- Browser calls a CGI script[2] or a computer program on the server and passes the data filled. The program saves the data and can act differently according to the data it gets.
- Browser sends the data through e-mail. Then a program accesses those mail messages and collects the data.

In general CGI script calls offer a lot of new possibilities to the surveyer, but on the other hand one should not forget the benefits of the second option. When message (with the responses) is mailed, the sender's e-mail address is usually known. That means his or her identity is revealed, what can raise some privacy issues.

2.2.1 Plain HTML forms

In this approach the questions in a questionnaire are ordered one after another on one single page. Let us review some of the key weaknesses of this mode: respondents are able to view the whole questionnaire, sometimes by mistake skip a whole screen and control questions can not be built in.

The interviews that work on the principle of the HTML forms are in essence just electronic versions of the mail interviews, which means that with this kind of interviews all limitations and recommendations for the mail interviews must be taken into consideration: they must be short, simple and without complex shifts.

Described kind of form-based Internet interviewing uses a simple CGI script, which takes the respondent's answers and stores them in a database. Sometimes, more capable CGI scripts that support checking of not answered questions and also allows for answers consistency checking of answers are used.

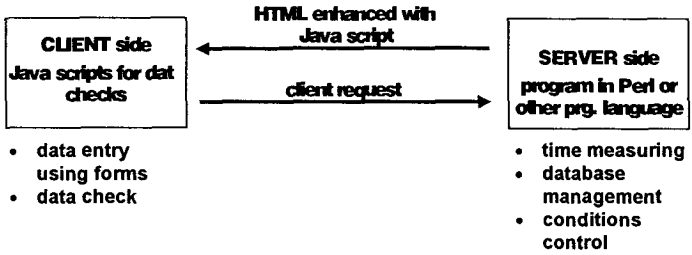
2.2.2 Computer Assisted Web Interviewing – CAWI

In this mode the questions (or groups of questions) are ordered one after another on several pages. This is the mode where CGI support at the beginning and at the end of each HTML form is used.

It includes all the possibilities of computer assisted interviewing required by sophisticated questionnaires: complex shifts, rotation and mixing of the questions, time measuring, but most important, all types of control of the received answers, control over the values that are allowed at a certain question, as well as the control over consistency of the answers.

CAWI is the Internet surveying mode that is most frequently used among professional research firms.

[2] CGI scripts are simple programs that enable interaction between HTML pages and WWW servers.



2.3 JAVA-applets-based WWW interviewing

With the introduction of Java, ActiveX and Javascript, almost everything can be done on the client side, without interacting with the WWW server. It is still a form of WWW interviewing which uses HTML only as a shell. The respondent's computer along with a WWW page downloads and executes a computer program in Java or ActiveX. The program has usually nothing to do with HTML forms. It acts as an independent program running in a WWW browser window.

Web interviews that use such an approach have basically no limitations about the data control, shifts, etc. Such method would enable the data control as one goes along, even though the questionnaire is layed out on one single page.

In practice it turned out (Java applets are no longer used by GVU) that these kinds of solutions are in principle extremely favourable, yet they are relatively slow, because the whole questionnaire (computer program) must be downloaded before the interview actually starts. The downloading procedure takes too long, especially when slow Internet acces on the responedent's side is used (ie. dial-up access). Therefore, such a mode is temporarilly abandoned and shifts are performed by consecutive HTML forms (CAWI). Also from the respondent's point of view, incompatibility with usual HTML-forms-based data entry should be mentioned.

2.4 E-mail interviewing with HTML support

Recently, software packages for the Internet communication (Netscape Communicator and Microsoft Explorer) were extended from WWW browsers to packages that include WWW browsers, e-mail, Internet conference tools, and HTML authoring tools.

With this improvement, e-mail messages can be in HTML format. This means that all features of WWW pages (e.g., pictures, tables, multimedia and forms) are supported. With the more intrusive nature of e-mail and ease of data entry using HTML forms, this Internet interviewing mode will be used in the future at least as a

new feature of e-mail invitation letters. For now, the problem is a large proportion of Internet users uses e-mail clients that do not support HTML.

3 The methodological issues of WWW interviewing

Two basic modes of performing interviews over the WWW have to be distinguished. The interview can be designed for general population that decides to participate on their own. Or, it is possible – and this method is by all means more promissing – to direct the interview to some well known target population. Such populations (with access to WWW) are at this time relatively rare, but increasing in number. Examples of small target populations are certain professional associations, some groups of firms and organisations. This method is especially suitable in the context of so-called mixed mode surveys – interviews where respondents have different types of interviewing modes (personal, telephone, Internet) available.

In the next section the discussion of the of the most important methodological issues of the WWW interviewing is given. Problems of complete WWW surveys, not just cases of simple voting (one or a few simple questions on the WWW pages), are discussed.

Several methodological issues of Internet interviewing were tested in RIS[3] national WWW surveys in the years 1996 and 1997. Both of them were conducted in April and May. Sample size for the year 1996 was 1200, in the year 1997 the number tripled to the size of 3500. Some of these issues are presented in this section.

3.1 Sampling

3.1.1 Self – selection interviews

In cases where respondents respond by their own initiative, the question of the sample being representative of a certain population becomes extremely important. A similar problem arises with ordinary mail interviews that are attached to some newspapers or magazines. The common denominator is self-selection: respondents decide (without any further incentive and without the personal influence of the interviewer) whether or not to participate. Only specific groups can be obtained and included into the sample, which makes generalization to the whole population impossible.

Experiences of other WWW surveys (mostly GVU[4]) show that the socio-demographic structure of the respondents is stable and therefore enables observation

[3] Research on the Internet in Slovenia, see http://www.ris.org

[4] GVU – Georgia Tech Research Corporation conducts the biggest international WWW survey twice a year, see http://www.cc.gatech.edu.

of trends. WWW surveys are specific in a sense that they include respondents who are more experienced and active Internet users. Other essential deviations are not perceived.

In marketing and media research only few surveys use probability samples. Nevertheless the results are generalized to the entire population, since variables are mostly robust enough for these kinds of deviations, if only a basic socio-demographic structure (quotas) is ensured. By weighting it is possible to adapt WWW surveys to the actual structure of the Internet users, but it doesn't really make a lot of sense. The purpose of the WWW surveys in current stage of the Internet use is first of all the reflection of the population that is actively using and responding to WWW environment.

We can influence the response rate of the WWW surveys by a method of advertising, rewarding, setting a time of conducting the WWW survey and by sending e-mails (but within certain socio-demographic segments). Time and place of responding are not really essential. By regulating the advertising, it is possible to influence the proportions between segments, but not the values of the answers within one segment. Since the data about the structure of the segments are usually available, the WWW survey is in the first place suitable for analyzing the most active populations on the WWW within specific segments. It is possible to do so because of the large samples that are obtained by such interviewing method. Frequencies and percentages on the whole WWW are illustrative and informative, but they can (because of unproportional number of younger respondents) include contents that are difficult to interpret.

3.1.2 Surveys based on a preselected sample

Most often a sample frame for WWW survey is set with the help of e-mail. On such basis the units are selected into a sample and an invitation to participate sent to them. In the frame of combined interviewing (mixed mode surveys) only those members of the target population with access to the Internet can be obtained.

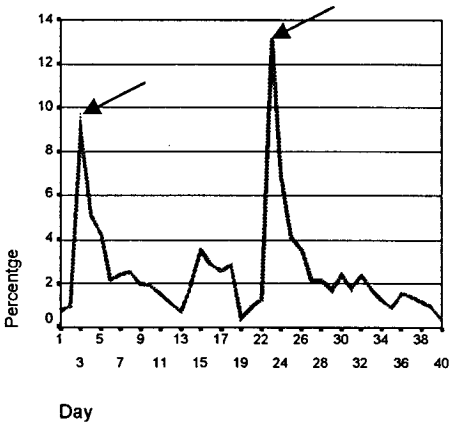
If we ignore the fact that in general only one third of the Internet users have their own e-mail address, the main inconvenience is mostly the absence of quality sample frames. There do exist more or less extensive mailing lists, which are found by specific computer programs (spider). Of course certain domains are often closed for them. Besides, such lists include organizations where many un-active users are listed. Special attention must be devoted to problems of multiple e-mail addresses of a single user (in the RIS survey almost one third of the respondents with an e-mail were in possesion of multiple e-mail addresses). There are also some respondents who would never disclose their e-mail address (according to RIS, that represents about 25% of Internet users).

Much more convenient is to form a sample frame in the case of a special population, for which we already have their addresses, e.g. students, professional

associations, webmaster populations, companies with intranet. In the case of high levels of the Internet usage among a target population, the interviewing over the Internet is the cheapest way of data collection, and it is user friendly (respondents decide by themselves when to participate).

3.2 Promotional activities

The method of promotion is very important especially in the case of WWW interviewing by self-selection. Let us illustrate this issue with the RIS '97 WWW survey where the time of the respondent's answering the questionnaire was coded. In the figure below the day of answering is presented. Two modes are seen at the 4th and 23th day of the survey.



Both days coincide with the promotional activity of sending notices by e-mail. The third day a message was sent to the respondents who stated their e-mail address in a previous RIS WWW survey; on the 22th day the notices about the survey were delivered to the respondents by Arnes (the biggest Internet public access provider Academic research network of Slovenia). The key observation is that direct promotion of WWW surveys by e-mail has a decisive influence on the response.

In both surveys respondents were asked where they had found the information about the survey. In the RIS '96 survey most of the respondents reached the questionnaire through the main Slovenian Internet directory (Mat'kurja), 19% were informed by login messages, 17% were informed by messages on other WWW pages, and 13% were informed by printed media.

The proportion of Mat'kurja was significantly reduced in the RIS '97 survey. There was an increase in other WWW pages and the proportion of users who received the e-mail invitation. Further analysis shows that the source of information about the WWW survey also influences the demographic profile of respondents.

In the RIS '97, the computer program for interviewing was improved in such a way that it kept records about which WWW page the respondent came from. In this way it was possible to measure the efficiency of advertising on WWW pages and find out what types of respondents came from certain WWW pages.

The conclusion is that the most important impact for the decision to participate or not related to e-mail promotion. Almost half of the respondents were obtained this way. The second most important factor was advertizing on the most visited pages.

3.3 Interaction with respondents

One of the key advantages of WWW surveys is the simple fact that it is very easy and almost without costs to perform follow-up communication. This also includes the possibilities of distributing the results of the survey.

WWW surveys not only attract attention, but also allow the researcher to keep in touch with the respondents, especially when the survey is repeated periodically. Both are based on retrieving feedback information about the survey and by building an e-mail database of respondents. At the end of the WWW questionnaire, the respondents are asked about comments and their e-mail addresses. In the RIS surveys, in exchange for their address, we promised the respondents the first results of the survey. Surprisingly, in RIS '96 as well as in RIS '97, approximately 2/3 of respondents left their e-mail addresses.

To each respondent was e-mailed a thank you letter and an URL address, where results were available. This letter was made for each respondent separately, with the exact duration time of each interview stated (for example "thank you for your 5 minutes and 35 seconds of your time for answering).

This impressed the respondents. The intention of the idea is to motivate the respondents to participate again next year. The key advantage of WWW surveys lies in a simple fact that feedback communication and additional reminding of the respondents is easy and almost cost-free. It also includes the distribution of the results of the survey.

3.4 Rewards

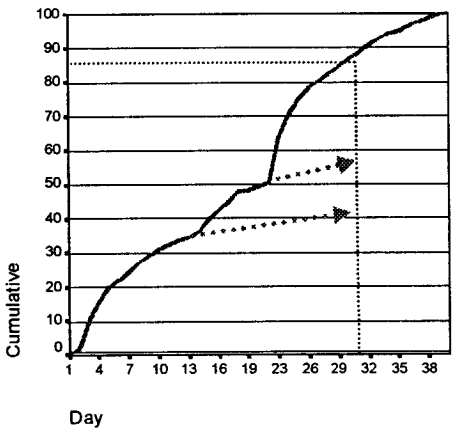
Rewarding the respondents for their participation in the survey undoubtedly contributes to a larger response. But it can also have negative effects: rewards mean additional costs, but most of all, the risk of over sampling specific types of respondents (that respond better to rewards). The effect of the rewards can also be

over-stimulative in the sense that some individuals answer the survey several times. The next question is in what way the rewards should be mentioned: at the beginning or at the end of the survey, or perhaps in some completely different place? In the GVU WWW survey, because of initially reduced responses they introduced rewards in the amount of 100$ which were to be received by randomly chosen respondents.

3.5 Time

By WWW surveys we are enabled (like in other forms of CADAC) to measure the beginning of the interview and the duration of the whole interview. Such possibilities allow us to control the process of interviews and to experiment with the length of the interviews.

Measuring the start of the interview is important also from the technical point of view. With analysis we can predict possible server overload that results in difficulties when accessing the survey. This is an extremely important issue when large scale (global) WWW surveys are conducted. For illustration let us review at what hour the respondents in the RIS '96 survey decided to participate. In '96 there were two modes, one between 10 and 11 AM (nearly 10% of all respondents) and the other one around 20 PM (around 6%). We had a similar distribution in '97.



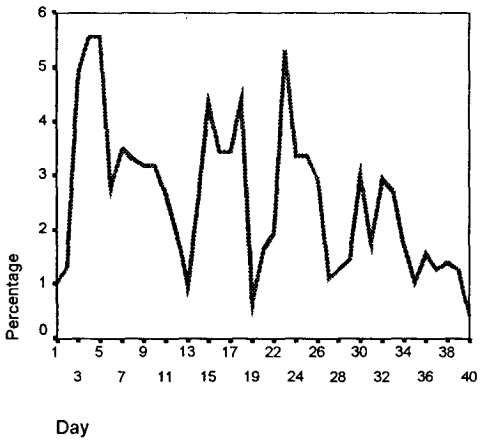
3.6 Duration of the survey

A period of one month became widely accepted to carry out WWW surveys where the sample is not preselected. The same was the case with RIS '96 survey. In RIS '97 this period was extended to 40 days.

From the figure from the previous page we can see that approximately 13% of the interviews were gathered in the last ten days. Therefore the duration of the survey does not have such an influence as additional incentive (promotion), since we can see that by every additional incentive the cumulative curve quickly flattens. The dotted arrows clearly illustrate the level we would reach in one month without the additional incentives.

The demographic structure of the respondents at the last days of the survey shows more women, more respondents above 40 than average and a higher proportion of new users. Thus by prolonging the survey less, intensive Internet users were attracted.

In the figure below, the day when the RIS '97 survey was responded to is presented. In this case users who received e-mail messages are excluded to remove the influences of e-mail notification.



If we remove the influence of the incentive in the form of e-mail the reduction in the numbers of answers considerably lowers. In the chart, the gaps can be seen. Those are the periods when number of the respondents was reduced, mostly on weekends. Despite the fact that weekends were "slow", we notice that the weekend respondents differ from others. They are not likely to be employed (they access the Internet from home), less often women and more often young respondents.

3.7 The length of the questionnaire

The WWW questionnaire survey in RIS '96 survey included standard Internet-usage questions (frequency, areas, and background). Altogether, there were 20 questions which on average took 7 minutes of respondent's time.

In the RIS survey, two different layouts were tested. The first one was the commonly used one - one long scrolling page for the whole questionnaire. The alternative one was a layout where each question block was put on its own page (CAWI); the next page appeared only when the previous was finished. The software automatically allocates the proper layout to each respondent.

The second layout has many advantages over the first one. When using the second one, it is very simple to implement the use of conditions in the questionnaire. Otherwise Java applets needs to be used which results in unnecessary waiting for questionnaire download, an issue discussed before. When using the second layout, the measurement of time used for each question block is done automatically.

The impact of the layout on the completion rate and length of the interview was also tested. Here, by the completion rate we understand the ratio between number of complete /questionnaires and the number of all attempted/started questionnaires. The results were calculated separately for text browsers (Lynx) and graphical browsers. The results are presented in Table 1.

Table 1: Mean of the responding time and the completion rate

		mean (sec.)	completion rate
Text	**multiple page**	553	76.1%
	one page	468	76.9%
Graphical	**multiple page**	466	83.5%
	one page	368	85.4%

It is evident that the completion rate is highly influenced by the type of browser used. The average length of interview is higher when text browsers were used and also higher when multiple page layouts are used. Also, in same environment (text or graphical) interviewing time was higher when multiple page mode was used. The most important result is that there are no statistically significant differences in completion rates which means that no significant differences occurred when text browsers with multiple page layout were used compared to single page layout. Therefore, there are benefits of the multiple page layout (CAWI) as discussed before, and there are no limitations for using it.

4 The technology used in RIS surveys

All RIS surveys mentioned above were created with the software that supported Integrated Computer-Assisted Data Collection (ICDC). The software that was developed enables easy transformation of the questionnaire to the surveys of different modes. The basic idea is to create the questionnaire once and let software create the final layouts for all modes of data collection.

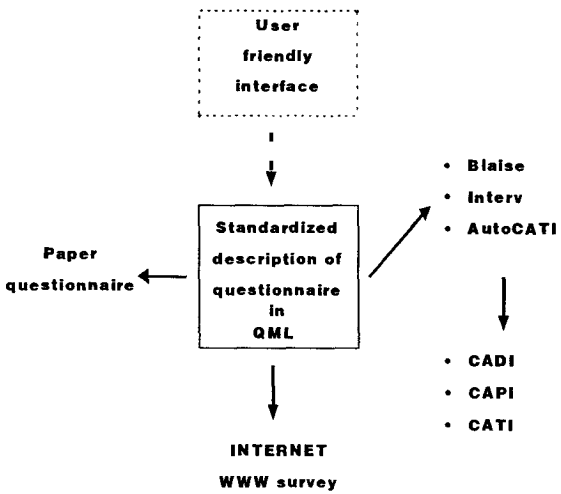
All questionnaires in RIS project were thus written in the questionnaire markup language (QML[5]), which is based on SGML (Standarized Generalized Markup Language) - an international standard for data description. The software than enables automatic conversion to CATI, CASI and CADI format which can be directly applied for interviewing and data entry.

The questionnaire can be also printed in the form that can be used directly for an e-mail survey. Of course, some additional design work may be needed on the layout.

The questionnaire can be also converted directly to the specific form of standard software such as Blaise and InterV. This may be extremely useful when cooperation is needed between agencies / companies using different software. A CATI / CADI / CASI software interpreter for QML has been also developed.

Another important output is the format for WWW surveys: HTML, Perl and Java scripts. The software automatically creates and designs a standardized layout of the page on WWW. The important point is that QML supports also HTML source code. That means that multimedia WWW surveys are also possible.

The above described software can be schematically expressed in the figure below.



[5] See http://www.cati.si/cati/qml for more information.

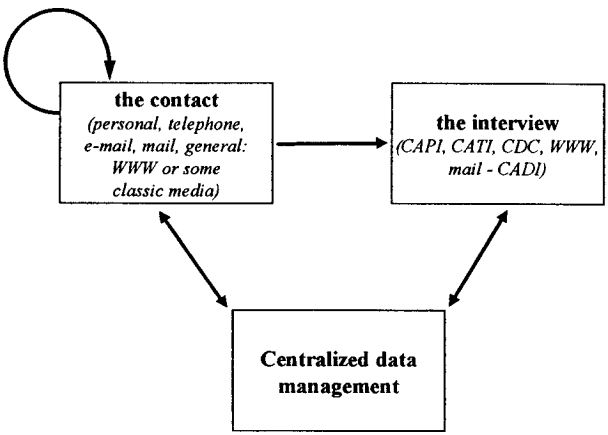
5 Integrated Computer Assisted Data Collection

The above-described technology enables considerable flexibility in designing the surveys. In general, the survey process can be split in two parts: the contact with the respondent and the interviewing process itself. Both components can of course be performed together; however, there are situations when the two components are separated.

The above scheme enables large flexibility in selecting the most suitable mode / technique to perform the survey. Especially, the integrated computer assisted data collection approach strongly supports the mixed mode surveys. The advantage is to give the respondent the comfort of selecting the preferred mode and also the time of the interview. Of course, this is reasonable only in the case of a relatively motivated target population.

In the majority of surveys with less salient topics, the aggressive approach based on face-to-face or telephone contact, which is immediately followed by the interview, may still be the preferred option and should be in no way neglected with availability offer of the mixed mode survey.

On the other side, the approach described above is especially suitable not only for mixed mode surveys but also for the collection of the administrative data.



6 Discussion

In this paper, the technological and the methodological side of using WWW surveys is discussed. It seems that surveying on the Internet will become extremely important in a few years. With greater percentage of population connected to the Internet, the representativeness of the obtained sample to the general population will become better. Therefore, the importance of this new method of surveying will grow.

From the technological point of view, it is important that all major commercial software applications for computer assisted data collection include WWW surveying modules. Even more, software developers are establishing their own services for WWW surveying. There are at least two reasons for this direction: the first one is completely materialistic one, software producers decided to run a service rather than just selling software applications, which adds value to their software. The second one is that in WWW surveys, the stability of the WWW server is one of the most important issues.

The central question about WWW surveying is the following one: Is it really so inexpensive? The answer depends on what kind of surveys is conducted. A stable leased line and stable hardware (WWW server) capable of serving many simultaneous respondents are needed. If only one survey is needed, initial costs are very high.

There are also other important issues in WWW surveying, especially if we are running surveys open for anyone. They are transparent and very vulnerable for several types of attacks regarding content of the survey and possible errors. Also, the public image of the surveying organization is an important issue when the survey tends to attract large populations.

The convergence of interviewing modes, what we call ICDC (Integrated Computer Assisted Data Collection), should be also mentioned. The key application that enables ICDC is platform independence of Internet applications supported by WWW browsers, servers and e-mail readers based on different operating systems (including handheld computers and palm-computers). From software developer's point of view, this offers a whole new perspective to the users. Applications for computer-assisted interviewing need to be programmed only once. The same interface can be now used for CATI, CAPI, CAWI; computers themselves can be located in the same place (call centers); they can be located on different places but linked together with a Wide-Area Network; or they can be linked together non-permanently with dial-up access.

All these new promising possibilities will bring new problems that will have to be discussed in the future.

References

[1] Batagelj, Z. (1996): *Raziskovanje (prek) Interneta.* DSI - Dnevi Slovenske Informatike '96, SDI, Ljubljana, 1996. (in Slovene)

[2] Kottler, R.E. (1997): *Exploiting the Research Potential of the World Wide Web.* Paper presented at Research '97, London.

[3] Pitkow, J.E. and Recker M.M. (1995): Using the Web as a survey tool: Results from the second WWW user survey. *Journal of Computer Networks and ISDN systems* **27.** (electronic document)

[4] Pirkow, J.E. and Kehoe, C. (1997): *GVU's 7th WWW User Survey.* (electronic document)

[5] RIS (1997): *RIS '97 po WWW - rezultati po podskupinah.* University of Ljubljana. (in Slovene)

[6] Saris, W.E. (1991): *Computer-Assisted Interviewing.* Series Quantitative Applications in the Social Sciences, London: Sage University.

[7] Vehovar, V. and Batagelj, Z. (1996): *Methodological Issues of WWW Surveys.* Paper presented at InterCASIC '96. San Antonio, http://www.ris.org/casic97.