NOT DROWNING BUT WAVING: RESEARCHING THE INTERNET

Nigel Jacklin, Objective Research
Peter Highland, Financial Times

Introduction

The stringent deadlines of the Worldwide Research Symposium meant that I had to write this paper in August and yet wait until October to give it. I haven't updated it except where there have significant changes in the data. In some ways this summarises the whole dilemma of Internet research. How can one research something in the traditional sense when it is constantly changing shape. The press get criticised by agencies for not producing more up-to-date figures on readership and circulation, even when the changes may not be that dramatic. How much more trenchant will their criticisms be when readership/viewership of an Internet site may have gone up by 50% or the site may be totally unrecognisable from the site that was researched in the first place.

I am not going to give you a wearisome history of our Web Site development. I know how boring other companies strategies and trumpeting of successes are. The key point from the research point of view is that users log by means of a short demographic questionnaire (the only price of entry). This covers details that most of you in this room could recite in your sleep. The site contains about 1/3 of the content of the printed product but there are the inevitable drill down facilities into some of our related databases such as Extel and FT Profile, and interactive opportunities with Louise Kehoe our California correspondent for example which are not available in the newspaper. 56% of users of the site are regular readers of the newspaper. So just over 40% of users seldom read or never read the FT.

One of them lives in Antarctica, proof if it were needed the Web site overcomes some of the delivery challenges an international newspaper might encounter. Perhaps most ground breaking is the deal we have recently struck with Microsoft which means that the FT icon will be one of admittedly 200 others to be displayed on millions of Microsoft browsers.

We sell advertising on our web site in a number of forms... banners, content specific hypertext links to other sites. The research data is primarily used to back up this sales effort.

The site started in May 1995 and the entry questionnaire was introduced as part of the relaunch in April 1996. Registered users have grown steadily from well... zero, to over 470,000 users in a year and half. User patterns vary enormously - there are people who log on and never come back; there are people who use us for 8 weeks curiously or so and never come back. About 5-10% are core users. There was even a man in Spain who dutifully fills in his questionnaire each time he logged on not realising that once is enough and there are those who have lost their passwords and therefore log on again under a different password. We mail these people back and have installed a password look up system to give users back their first password, in order to keep the subscriber list as accurate as possible. Interment users can be impatient, intolerant of this research based entry device. Here's a quote from someone from the US which highlights the issue.

'Have you seen the detailed personal information the FT asks of those who want to read the FT on-line? I don't have to tell my newspaper vendor my annual income, age, gender, sex...'. Not sure what the difference between gender and sex is. But you get the message. She's sore.

On the other hand, they are often enthusiastic communicators who respond with exemplary speed to the special research requests we have run. A quarter of respondents to a recent survey were recruited by e-mail, responded to the mail, visited the special web site and filled in the survey within 12 hours.

Overall however, it's complicated and sometimes the experience we have with researching the printed medium is scant help. Not drowning but waving? Yes, but it's a close run thing.

Research Challenges

You can tell I've been on a positive thinking course. Challenges, never problems. There are four of these essentially. What are the key differences between researching FT.com and the Financial Times newspaper? There are four of these essentially.

1. The fact that the information is generated automatically and held on our own servers. Unlike the newspaper business where research is painstakingly gathered by contacting the end user and research companies agonise about how they can get more people to spend 40 minutes discussing their personal finance with some one they've never met before. Here the problem is limitless data. And it keeps coming. So the research challenges for me and for research companies are not to construct and do the research (except perhaps at the questionnaire stage) but to help me swim in oceans of data, throw me a life jacket of interpretation and train me to catch something edible out of the water and package it into bite size chunks for editorial and advertising use from a limitless choice of options. I think this metaphor has now run its course.

2. The fact that this information relates to every detail of every customers use of the site. It is a census (not a sample from which usage patterns have to be modelled using recency and frequency data and various other assumptions). It's the whole picture and probably more than the customer wants or can absorb. In practice, advertisers seem to be happy with page views per month (all in all pretty blunt measure) and may continue to be so until money invested in the medium justifies more complicated measures. In time we will install ad management systems which allow advertisers to access the data themselves rather than rely on us to give it to them.

3. That there is a large volume of this data constantly flowing into our computer; the information is up-to-date and large enough to make processing capacity an issue (a thing of the past for most readership surveys). Nine months data (once compressed) takes up 3 Gigabytes, nearly 300 times the size of the UK's National Readership Survey data held on the IMS system.

   All in all, I don't need a research company. I need a data processor with imagination as well as one with a bit of customer focused intelligence (I'm trying not to use the word strategic here), with a big powerful machine within easy reach.

4. Advertisers require information about the actual performance of their campaigns, which we are able to provide. The age old agonies that press underwent to prove that people saw ads in print, reacted to them and behaved differently is solved at a stroke. We know what pages people look at and at what time. The upshot of this 100% accurate research on demand is possibly payment by results but we have resisted this so far and, in line with the Internet Advertising Bureau (IAB) guidelines, have adopted page view as a standard measure (rather than click through). We then give the advertiser the % of clicks that turned into page views plus all the demographic breakdowns.

Another side issue to plague the researcher is that the server logs an impression even though the user may have switched the graphics (and therefore the advertising) off in order to down load the data only. They will still be logged as a viewer. However on the plus side, the server does not measure so called 'cached views' - repeated viewings of the same screen in the same session, which therefore generate more page views than are measured. We tackled this issue in our first on-line user satisfaction survey which indicated that 85% look at the full graphics, but there is more work to do on accounting for both these anomalies.

The precise information requirements of advertisers and agencies are still developing as the medium grows in importance and the media and advertising community learn more about it, but whereas the printed media are measured on joint industry surveys, these have not yet developed to a serious extent in the advertising markets in which web sites operate. We have adapted our site to adopt the IAB guidelines in terms of terminology and ad sizes that web sites are increasingly using in the US to give some consistency of measurement, but initiatives to bring this together so that advertisers can compare the performance of sites has proved slow. Lack of compatible demographic data is one issue. Non existent demographic data is another.

All these issues are not solved by the use of conventional questionnaire based market research, but do require the broader skills of the researcher in terms of data analysis and understanding customers and markets. To summarise here are the four big research challenges:

- the need to provide a base level of information
- the need to handle a large and constant flow of data
- understanding customer needs and usage patterns
- the desire for compatible data.

Let's go through what we have done so far on each of these.

Base Level Information

We provide information both pre and post sale data. In this respect the medium is closer to television than to print. A typical presentation to a potential advertiser in August of this year included the following information:

- 480,000 registered users
- 14,000 average daily users
- 60,000 users in past 7 days

- 46% of registered users are aged under 35
- 37% resident in UK, 24% in US, 4% in Canada (the third largest country)
- 20% earn $80,000 or more
- 22% control expenditure budgets in excess of £100,000.

Nothing too earth shattering there. Pretty traditional historic style measures. However any variances to this profile can be given to the advertiser after their ad runs. In order to help us plan which pages to include the banner on, we also look at the profile of the users of the different pages.

In addition to this base level information we have now started to carry out our own research. To date we have carried out two major projects, the first to gauge user satisfaction, the second to provide personal finance advertisers with more information. The user satisfaction survey was presented as a forced screen to a random selection of users when they logged in to the site, whilst for the personal finance survey we e-mailed a sample of recent users and directed them to a web site where they could fill in the questionnaire. Whilst the forced page questionnaire generated only a 15% response rate, the e-mail survey generated 23%, low compared to most of the industry surveys discussed here at the Symposium but comparable to many single wave postal surveys. E-mail is a method we will use again, perhaps considering a reminder in order to enhance response. The advantages it has in reaching some respondents were illustrated by one of the winners in the prize draw (used as an incentive for the personal finance survey) who picked up his e-mail on an offshore oil rig in the Persian Gulf, and sent it back using a combination of land wires and satellites. This type of respondent is likely to be under-represented in conventional research, never mind conventional media. Probably quite a lot of disposable income too. And time to think what to do with it.

Both surveys were conducted in house; the main difficulty we encountered was at the analysis stage, our software not having been designed to do cross tabulations and weight data. This has, therefore, restricted the extent to which we have been able to use the data and this is where the big outside data processor comes in.

Handling the Data

We have three main sets of information which we keep:

- a registration file, currently containing records for 480,000 individuals
- a daily log file, recording the daily use patterns
- a file to interpret the page records in the daily log file (i.e. a sort of code book).

Each day we have 15,000 users who visit, on average, 8 pages... i.e. 118,000 page views. Whilst, in web page terms, each page may have 10 or more components (such as individual graphics, text or advertising banners), each of which is recorded separately, we strip this data down and store only full pages to reduce storage. Initially we generate over 3 million records each day comprising a date and time stamp, user i-d and page/component i-d. This file is reduced each day, before being saved, as we are only interested in which pages they have visited and which advertising banners they have seen or clicked on. I mentioned that 9 months worth amounted to 3 GigaBytes. Processing this amount of data can take three days on our current system. This kind of volume is a serious barrier to entry for many potential data analysis supplier.

Frequency of reporting, therefore, is theoretically possible on an up-to-the second basis. However, our more pragmatic approach is to monitor the number of registrations on a daily basis, together with the number of users each day. The main point is that advertisers can get their effectiveness reports whenever they want them and they will always be up to date. So a big advance on the world of print.

Taking a more philosophical viewpoint, the fact that 'anything is possible' makes it more difficult to figure out ‘exactly what are we going to do’. The need to focus our analysis of the data is paramount if we are to avoid drowning in numbers. For management purposes, therefore, given that the site is advertising funded, our decision has been to set a few basic standards, such as the need to maximise the number of people who use the site at least a few times a month, rather than the number of regular users. This allows us to maximise what we deliver to our advertisers in terms of coverage.

Understanding Customer Needs and Usage Patterns

We are embarking on a major new project, with the help of an external analysis bureau, which will provide us with a better picture of peoples usage patterns and the extent to which different customer groups display these patterns. At this stage the work does not involve any of the ‘questionnaire’ type of market research, being carried out by a bureau which specialises in analysis of customer databases. Specifically we will be looking to see whether use tends to be needs driven (resulting in bursts of use for a particular period) or habit driven (splitting into traditional regular and occasional use). We will also be trying to unravel the different stages of use, from trial, to establishing use patterns, and (hopefully not) lapsing. This will allow us to target high advertising value users who are potentially about to lapse, and encourage them to continue using FT.com via e-mail, developing the dialogue aspect of the Internet. The work will draw on both the registration data and use data, both on a daily basis and in terms of type of page, the kind of research which print media editors can only dream of on such a large scale.

One practical hurdle in starting this work has been the scale of the 9 months database. Here, in order to be able to download the data we have had to take a 1-in-10 sample. As the daily log file is held in page number (rather than user number) order, taking such a sample of users rather than pages initially looked impossible, however, by pulling the page records for all users with an i-d number ending in 1, we were able to generate a perfect random sample of a more manageable size.

From a customer database point of view, FT.com is a database marketers dream. On a practical level this is because (at the moment) all data is held in a common format and is complete. In terms of richness of data, it has more detailed data about what customers have come to us for, rather than simply recording what we have sent out to them.

The results of this work will be available at the time of the symposium.

The Desire for Compatible Data

It is perhaps unsurprising that, in such a new and fast moving medium, there is a lack of compatible data. As a leading advertising funded site, therefore, our aim is to keep at the forefront of what advertisers and agencies require and ensure that we provide data to help them both plan and evaluate their advertising.

There are some initiatives for joint information, but they have yet to gain much momentum. There are a number of reasons for this:

- some of the auditors, such as ABC, BPA and IPRO audits, offer lowest common denomination data, at a significantly lower level than we already provide. However they are expensive and tend to overclaim their dominance of the business. The web business needs to generate its compatible data from within rather than be monitored from outside.

- the number of sites with whom we compete for business on a serious level is fairly limited; whilst there are larger more consumer orientated sites, it would be foolish for us to aim to be compatible with them... the first step is to define which sites make up the market in which we compete

- once this stage has been reached, the need is to both agree a common measure of advertising exposure as well as standard demographics/marketing definitions.

Whilst the market itself is taking care of the advertising exposure measure (we have moved from nothing to hits and now to page views or impressions) the area of common demographics is more difficult. Site owners would need to reclassify their user databases and registration forms to a common standard. This may, however, be no bad thing, as it would also mean clearing out a lot of the dead wood. Once this was done, it would be perfectly feasible to create an industry database by merging the files of a number of sites, identifying common users by their e-mail addresses. This works fine, so long as only one person accesses each site using each user i-d (i.e. no looking on anyone else's PC) and that no-one has more than one e-mail address. As both of these assumptions are invalid, there may be a case for doing some interviews! They may even be more palatable than e mail.

Conclusions

As a medium, FT.com differs significantly from the newspaper both in terms of how 'readers' and advertisers use it.

Our commercial relationship with advertisers is different, more accountable. As they put more money into the medium, their research demands will increase.

We do, however, have vastly more data on use of our own site, but at this stage lack much data on which other sites our readers/viewers go to.

The analysis of the data presents different challenges, mainly due to the large volume of it and the fact that it is constantly updated... in a land where everything is possible, it is even more difficult to decide what to do.

That sounds like the world wide web to me.