Industry Canada

Music File Sharing Study 2006

Methodology Report

August 31, 2006

Registration number: POR-335-05
Contract number: U5500-050472

Proprietary Warning

The information contained herein is proprietary to Industry Canada and may not be used, reproduced or disclosed to others except as specifically permitted in writing by the originator of the information. The recipient of this information, by its retention and use, agrees to protect the same and the information contained therein from loss, theft or compromise. Any material or information provided by Industry Canada and all data collected by Decima will be treated as confidential by Decima and will be stored securely while on Decima's premise (adhering to industry standards and applicable laws).

Toronto
2345 Yonge Street
Suite 405
Toronto, Ontario
M4P 2E5

t: (416) 962-2013
f: (416) 962-0505

Ottawa
160 Elgin Street
Suite 1820
Ottawa, Ontario
K2P 2P7

t: (613) 230-2200
f: (613) 230-9048

Montreal
1080 Beaver Hall Hill
Suite 400
Montreal, Quebec
H2Z 1S8

t: (514) 288-0037
f: (514) 288-0138

www.decima.com

info@decima.com

Decima Research Inc. is ISO 9001:2000 Certified

Decima Research

Table of Contents

Introduction

Background

The widespread phenomenon of music file sharing over the Internet began in 1999, with the creation of the Napster program. This was a software program that could be downloaded onto a personal computer from the Internet free of charge, allowing users to exchange mp3 format audio files. Napster was the very first peer-to-peer (P2P) computer software programs, requiring a central server to connect users. The current generation of P2P networks does not operate through a central server, as each user establishes contact with others directly.

There is a growing literature on music file sharing and its impact on pre-recorded music sales. However, none of the existing studies have analyzed Canadian data, let alone focused on the Canadian case. The primary reason for this is the absence of Canadian data.

An effective means of collecting data on the variables of interest is through a survey of Canadians. Ultimately, this survey will help Industry Canada derive accurate and statistically reliable metrics around the following:

- Counts of the number of files shared over the selected time period;
- Purchases of pre-recorded music products (PMPs) over the selected time period;
- Number of Canadians who purchase PMPs; and
- Number of Canadians who engage in music file sharing activity.

Purpose

The purpose of this document is to detail the methodological considerations used in collecting this data. It clearly outlines assumptions used, procedures followed, and why certain decisions were made in mapping this data.

Methodology

Overview

This survey was conducted using a traditional telephone survey data collection methodology (CATI – Computer Assisted Telephone Interviewing). Specific dimensions of the research included:

> **2,100 telephone interviews with Canadian households over the period spanning April 19th through May 20th;**
>
> **National representation with special consideration by age, gender and region;**
>
> **Interviews were segmented based on whether or not the respondent had downloaded free music from a P2P (peer-to-peer) network during 2005;**
>
> **Quota-based, weighted distribution;**
>
> **Average survey length of 10 minutes;**
>
> **Interviews were conducted with respondents 15 years of age and older;**
>
> **Respondents were informed that the survey was being conducted by Decima Research on behalf of Industry Canada (i.e. sponsorship revealed study).**

Sampling Distribution

A stratified, quota-based sampling approach was used, since this generates substantive estimates across and within specific key segments of interest, which in turn permits extrapolation to the broader population with greater confidence. Previous research suggests that music purchasing and file sharing behaviours tend to differ according to a person's mother tongue, gender and age. Quotas were therefore set by controlling for age, gender and region, where region is used as a proxy for mother tongue – that is, Quebec is treated as Francophone. Furthermore, given the importance of understanding behaviours among both music downloaders and non-downloaders, quotas were introduced to control for this as well.

The grid below summarizes the final sample distribution across the quota cells. Industry Canada requested 1,000 responses for both downloader and non-downloader segments and approximately 100 responses per quota cell in order to meet desired standards of statistical reliability.

Region	Age/ Gender	Downloaders	Non-downloaders	Total
Quebec	Male <25	160	125	285
	Male 25+	90	125	215
	Female <25	160	125	285
	Female 25+	90	131	221
Total Quebec		500	506	1006
Rest of Canada	Male <25	160	132	292
	Male 25+	90	153	243
	Female <25	160	133	293
	Female 25+	90	176	266
Total Rest of Canada		500	594	1094
Total		1000	1100	2100

The maximum margins of error associated with the above segmentation are as follows:

Population Segment	Sample size	Margin of error
Downloaders or non-downloaders	N=~1000	+/- 3.1%
Quebec or Rest of Canada	N=~1000	+/- 3.1%
Downloaders or non-downloaders in each region	N=~500	+/- 4.4%
Males or females under 25 in each region	N=~285	+/- 5.8%
Males or females over 24 in each region	N=~215	+/- 6.7%
Non-downloaders within each age, region and gender segment	N=~125	+/- 8.8%
Downloaders within each age and region segment, under 25	N=160	+/- 7.8%
Downloaders within each age and region segment, over 24	N=90	+/- 10.3%

Research Material Design

The questions for the survey were designed in conjunction with the Industry Canada team, using the following iterative process:

- Industry Canada initially provided Decima Research with a draft questionnaire for review. Decima then provided recommendations regarding language used, survey flow, skip patterns, programming and interviewing instructions. An updated draft version was subsequently returned to Industry Canada, who then made modifications accordingly. Once all parties were satisfied that the questionnaire was ready to be pre-tested, the survey was programmed using CATI software.

Pre-testing

A pre-test was performed with 23 English and 23 French respondents. The general purpose of this pre-test was to evaluate respondents' comprehension of the questions, the completeness of the response categories, and overall response to the survey (…is there anything that might make respondents uncomfortable or angry? …are certain questions generating an abnormally high amount of non-response?). Further, two versions of the draft questionnaire were tested – "Version A", which asked respondents to estimate the total number of items purchased over a specified period (e.g. number of CDs purchased in 2005) and "Version B", which asked respondents to specify the total amount spent (e.g. amount spent on CDs in 2005). Respondents were randomly selected to receive either "Version A" or "Version B".

The distribution of surveys by version and by language is as follows:

	Version A	**Version B**
English	10	13
French	12	11
Total	**22**	**24**

Following the pre-test, Decima provided a summary of the results and recommendations for modification to the questionnaire. Among other issues, the pre-test revealed that 'Version A' was easier for respondents to answer. Based on the pre-test results, a number of modifications were made to the questionnaire.

Sample Selection and Contact

Sample Selection

Decima created the sampling frame using Canada Survey Sampler (CSS). CSS is a proprietary selection engine specifically designed to generate a random sample of telephone numbers to be dialed. The CSS maintains a comprehensive list of all populated exchanges across Canada, and is updated on a regular basis.

The CSS works by randomly generating 4-digit suffixes for these exchanges. These suffixes are generated in proportion to the percent population of the individual exchanges (i.e. a 90% populated exchange would experience twice as many 'hits' as a 45% populated exchange). As each suffix is generated, it is compared to the database of existing known phone numbers. If it matches a listed phone number, it is placed into the 'valid number' file. If not, it is placed in the 'orphan' file. Decima uses the valid number file as its primary calling list. This list is then supplemented with numbers from the orphan list. As with the random generation above, numbers are chosen from the orphan list in proportion to the percent population of the exchanges.

The supplementing of the sample file with the orphan file is intended to ensure that our sample scheme accurately emulates what is know as a 'Waksberg' RDD design, but is more efficient because it more effectively includes unlisted and new telephone numbers while significantly reducing the number of ineligible "not in service", fax and cell phone numbers that would normally be encountered with conventional RDD methods.

Respondent Selection

Because interviews tend to be completed by an adult in the household (notably those 20 years of age or older), once a household was contacted, the respondent was asked whether there was anyone in the household between the ages of 15 and 19 who could complete the survey, to ensure sufficient representation of this younger cohort within the youth segment.

Survey Administration

All interviewing was conducted through Decima's field division, Opinion Search, which operates one of the largest and most sophisticated research data collection operations in Canada. From its offices in Ottawa, Toronto and Montreal, Decima utilizes 325 state-of-the-art computer-assisted-telephone-interviewing (CATI) stations, which are fully networked and utilize sophisticated VOXCO technology that is among the most powerful on the market today.

Interviewing was conducted in English and French, based on the expressed preference of the respondent. Up to eight call-backs were made to each eligible or potentially eligible household during the interview period before classifying it as "not available." Fifteen percent of all calls were monitored by the supervisory staff. Frequency checks were administered throughout the data collection period to ensure that all questions were correctly programmed.

Sample Management

Sample was allocated in small segments and exhausted before new sample was substituted. All contact dispositions were tracked and recorded in the sample database. In many instances, the system automatically allocated sample records in response to recorded disposition. For example, the system automatically rescheduled callbacks for "no answers" and "busy" dispositions in a pre-set pattern to compensate for potential non-response on the occasions when individuals were not at home or their lines busy. For example, unanswered calls in the early evening could be scheduled for another day later in the evening; a weekend call could be scheduled for weekday evening and vice versa. For busy line signals, the schedule was created to re-attempt the record within a set time frame that same day. This entire process is consistent with the PWGSC standing offer for public opinion research.

Sample Weighting

Decima uses well-defined procedures for calculating weighting factors, based on established methodological standards and extensive experience in sample weighting on over literally hundreds of projects. This procedure involves calculating the actual population within each quota segment (region, age group, gender and downloader status), and the true proportion of the sample that would fall into each segment if the survey were conducted in

strictly a random basis (based on the most current population statistics available from Statistics Canada). Into this number is divided the actual segment sub-sample to produce a weighting factor that is then used to "weight" the data for that segment. While there are various ways of accomplishing this task, this procedure is the most straightforward and effective.

In this instance, the actual population of downloaders vs. non-downloaders was not known (or was not considered reliable). Thus, these proportions were estimated based on how sampling "naturally" occurred at the beginning of the survey period (i.e. up to the last day before any quota cells were closed). Details of this process are outlined in two steps.

Step 1. First, the true population proportions are determined by taking into account age, gender, region and downloader status, since these are the variables on which the quotas were based.

Using Statistics Canada Census 2001 data, we know the population proportions by age, gender and region, as seen below:

	Canada	Quebec	Atl	ON	West
Male Under 25	2,034,430	482,000	154,505	754,555	643,355
Male 25+	9,741,425	2,391,760	747,850	3,677,835	2,924,095
Female Under 25	1,974,705	467,480	153,125	733,275	620,800
Female 25+	10,530,965	2,604,635	819,730	4,011,665	3,094,940

For the downloader status, we look at the quota cells in the earlier days of the study before any of the quota cells were closed, since this is our best indicator of how the proportions naturally fall out when random calling is used. This distribution is seen below, and is based on a sample size of n=691 overall:

Type	Quota	Completed	Distribution
<25, Downloader, Male	320	50	7.2%
25+, Downloader, Male	180	64	9.3%
<25, Non-Downloader, Male	250	42	6.1%
25+, Non-Downloader, Male	250	176	25.5%
<25, Downloader, Female	320	31	4.5%
25+, Downloader, Female	180	53	7.7%
<25, Non-Downloader, Female	250	51	7.4%
25+, Non-Downloader, Female	250	224	32.4%

We assumed that the distribution within each region would follow the same pattern (because the overall sample size was relatively low, it was not feasible to look at the natural distribution within each region). Based on the national distribution, we calculated the population distribution within each region. Calculations for Quebec are provided in the table below.

Quota (Quebec)	Distribution	Population
<25, Downloader, Male	7%	430237
25+, Downloader, Male	9%	550703
<25, Non-Downloader, Male	6%	361399
25+, Non-Downloader, Male	25%	1514434
<25, Downloader, Female	4%	266747
25+, Downloader, Female	8%	456051
<25, Non-Downloader, Female	7%	438842
25+, Non-Downloader, Female	32%	1927462

= 7% * 5,945,875 (proportion DL * total Quebec)

For the final proportions, we needed to make them representative of the entire targeted Canadian population.

Quota (Quebec)	Population	Canada Population	Proportion within Canada
<25, Downloader, Male	430237	24,281,525	2%
25+, Downloader, Male	550703	24,281,525	2%
<25, Non-Downloader, Male	361399	24,281,525	1%
25+, Non-Downloader, Male	1514434	24,281,525	6%
<25, Downloader, Female	266747	24,281,525	1%
25+, Downloader, Female	456051	24,281,525	2%
<25, Non-Downloader, Female	438842	24,281,525	2%
25+, Non-Downloader, Female	1927462	24,281,525	8%

Step 2: Finally, we looked at our true sample numbers for each group within each province. For example, we collected 160 <25 DL Male in Quebec. Out of a total sample of 2,100, this sample proportion is 7.6%. Since this is higher than the true proportion (i.e. 2%), respondents falling into this profile need to be "weighted down". This is calculated by taking the true proportion within Canada (2%) and dividing it by the sample proportion (7.6%) to get a weight for "Quebec Downloading Males Less than 25 Years Old" of 0.233. This same procedure is used for each quota cell.

Follow-Up Qualitative Analysis (Post-Hoc)

Quotas for this study were based on a set of questions asked at the beginning of the survey to categorize the respondents into their respective quota cells. In order to keep survey costs at a minimum, calls made to individuals that fall into a "filled" quota cell are terminated as we continue to target the harder-to-reach quota cells.

A general set of questions related to music behaviour are asked upfront, which are then used for skip logic so that, for example, a respondent who did not purchase any CDs in 2005 will not be asked questions related to CD purchasing behaviour later on in the survey.

Similarly, our downloader screening question was asked in this early section of our survey, where we ask the respondent, "Did you download music from P2P (person-to-person) networks in 2005?", where the respondent had the option of giving a "yes" or "no" answer. Those who answered "yes" were asked questions in "Section 4: The Effects of Unauthorized mp3 Downloading".

Question 4.1b of the survey asks, "How many free music tracks did you download from P2P networks in an average month in 2005?" Normally, it is expected that 1%-3% of respondents arrive at this section, and then give a "don't know" or non-behavioural response (i.e. zero downloads). In this instance, 246 of our 1,000 quota gave a "zero" response or answered "don't know". Given the magnitude of this proportion of respondents, additional analysis was warranted to better understand true downloading behaviour. Post-hoc analyses were conducted to determine if these individuals should be categorized as downloaders or non-downloaders.

"Don't Know" Responses

Respondents who answered "don't know" are deemed to be downloaders for two main reasons:

1. Direct comparisons with the downloader group (i.e. those who gave a valid numeric response to Q4.1b) revealed that profiles of the two groups were similar across a variety of demographics, including age, student status, high interest in music, and proficiency with using the Internet; and

2. Response patterns among those who answered "don't know" revealed that this response was common to these respondents

throughout the survey, more specifically with those questions that asked for a numeric response.

"Zero" Responses

The more complex group to profile were those who answered "zero" to Q4.1b. We decided to go back into field and target those respondents who answered "zero" to Q4.1b. The following question was asked:

> *"In the survey, a number of Canadians indicated that during 2005 they downloaded free music from P2P file sharing networks, like Kazaa, LimeWire or Archambeault. But then when we asked them to specify the number of free music tracks they downloaded in an average month during 2005, many individuals answered 'zero'. So we're calling some Canadians back to try and understand this disconnect. We have identified a variety of possible reasons. Let me know if any of these apply to you. Is it because…"*

In this qualitative exercise, twenty-five respondents were targeted over the course of two evenings, and in answering this question, the intention was to help us understand how these individuals should be categorized for the analysis – that is, as downloaders or non-downloaders.

The reasons for the 'zero' responses can be divided into three main areas: they download sporadically, and not on a regular, monthly basis; they used to download from P2P networks but do not any longer and they do not download from P2P networks at all. Another less common reason is that they download few songs per year (i.e. less than one per month).

Based on these qualitative findings, we decided to assign these respondents a "downloader" status as well, with the subtle distinction being that they are not "regular" downloaders.

Appendix: Call Disposition Report

The chart below identifies the number of total attempted calls and the nature of these calls.

A (1-14)	**Total Attempted**	**44186**
1	Not in service	4508
2	Fax	711
3	Invalid #/Wrong#	1496
B (4-14)	**Total Eligible**	**37471**
4	Busy	273
5	Answering machine	6327
6	No answer	4846
7	Language barrier	855
8	Ill/Incapable	376
9	Eligible not available/Callback	2032
C (10-14)	**Total Asked**	**22762**
10	Household/Company Refusal	6516
11	Respondent Refusal	10369
12	Qualified Termination	186
D (13-14)	**Co-operative Contact**	**5691**
13	Not Qualified	3591
14	Completed Interview	2100
	REFUSAL RATE	**75.00**
	(10+11+12) / C	
	RESPONSE RATE	**15.19**
	D (13-14) / B (4-14)	
	INCIDENCE*	**38.90**
	[(14+12) / (13+14+12)]*100	
	[(CI+QualTM)/(NQ+CI+QualTM)]*100	