Chapter Three: Methods

3.1 Introduction

Three separate strands of enquiry were identified that were relevant to the study of the retrieval of online bittorrent information – the websites, the user-created metadata and the users themselves. Different methods were used for the study of each of these three strands. While these methods could be approximated to the three objectives stated in the introduction, each was designed to produce results that would be relevant to all of the objectives. In particular, the methods looking at the user-created metadata and the users were developed in tandem so that their results could be triangulated to better evaluate how effectively users' IR needs were met. Each of the three methods is detailed here in turn.

3.2 Review of bittorrent website features

The website strand of enquiry would be primarily exploratory in nature and would review the features available on a number of different websites. This method was developed and carried out first so that the results obtained from it could be used in developing the other two methods and locating sites for obtaining samples.

Design of the review

The review was to take the form of a table that provided an overview of the websites' characteristics. This was adapted from an evaluation methodology of web search engines developed by Chu and Rosenthal (1996) although rather than listing performance statistics it instead focused on the websites' structures and available features. Initially the only criteria were that the different fields within the torrent records of each website should be listed so it could be used for the content analysis method to follow. Additional features for comparison

were identified as the websites were studied, then consolidated before recording the data for each in a thorough and methodical fashion. They are listed prior to the results themselves in section four.

Selection of websites

By looking at a number of bittorrent websites three types of sites could be discerned: larger sites that dealt with many types of material, specialist sites that dealt with a single type of material and those sites that 'trawled' other bittorrent sites, using external data to create indexes of torrent records. It was decided that a number of sites from each of these types would be selected to ensure as wide a variety a possible.

A total of thirteen sites were selected for the review; five large multi-category websites, five specialized websites and three sites that harvested external data. The five larger websites and the three harvesting sites were chosen by virtue of having the highest traffic rankings amongst their peers as estimated by the website www.alexa.com. The five specialized websites on the other hand, were chosen to represent a diverse range of material.

3.3 Content analysis of user-created metadata

The method used to study the user-created metadata was the content analysis of a series of torrent records. This unobtrusive method was deemed appropriate in that due to the fact that the users creating the metadata were also possibly distributing illegal material. Any approaches made to users about specific content would likely be met with some suspicion, potentially affecting the quantity and quality of data received. It was also publically available and a relatively small investment in time and resources could yield a large sample size to work with.

Sample selection

Following the results of the review of bittorrent website features, the website www.mininova.org was chosen for the retrieval of bittorrent records. This was because the format of the records used by the website was typical of those found on bittorrent websites, with user-generated metadata being limited to three fields: the name, description and choice of category. In order to obtain records that were representative of all of the torrents uploaded to the site, the sample consisted of every record created by a user in a twenty-four hour period.

Coding the information

Once the torrent records had been successfully obtained, one hundred and fifty of them were randomly selected and put in a test group. This group was then coded by the researcher for the presence of metadata elements. Rather than impose an existing set of metadata elements (such as Dublin Core) on the data obtained, a grounded theory approach was adopted whereby elements were identified as they emerged from the data looked at (Strauss 1987). If an item of metadata did not fit into any existing category then a new category was established to enable its entry.

Core categories were not identified, but instead after the test group had been coded, the metadata elements were examined separately. High occurring elements were broken into separate components where there existed discrete categories and elements that occurred very rarely were either consolidated or discarded altogether. In this way, a comprehensive and manageable set of metadata elements was drawn up for the coding of the rest of the data.

As the data was coded there were several instances of metadata that was initially unrecognisable, usually technical abbreviations, references to other bittorrent websites or release groups. In these cases, the metadata item was queried using online searches. If this

did not clarify the nature of the metadata, an educated guess was made using other occurrences of the term(s) recorded online as a basis for categorisation.

Analysis of data

Once all of the torrent records had been coded, the data was stored in a spreadsheet and analyzed using the Statistical Package for the Social Sciences (SPSS) software. Groups within the data were identified based on the medium of the material described by the torrent record: audio, video and software. Together, these groups accounted for the vast majority of the total data. This grouping was done so that metadata elements that were exclusive to one of these groups could be analyzed separately and not be affected by torrent records where the element simply would not be relevant.

3.4 User Questionnaire

To effectively assess the information retrieval needs and methods used by the bittorrent community a self-completed questionnaire was designed and distributed to bittorrent users. The benefits of using such a method are the ease with which potentially large amounts of data can be collected and coded in a relatively short time. Conversely the risks are that there is 'little or no check on the honesty or seriousness of responses' and also potentially a very small amount of data collected (Robson 1993). In this case it was chosen because it seemed the least intrusive way in which to obtain answers to specific questions from bittorrent users and, as previously noted, therefore more likely to succeed with a naturally suspicious group of people. Also, as all bittorrent users have internet access, the questionnaire could be made available online. This enabled respondents to remain anonymous whilst also increasing the questionnaire's capability to reach large number of users. The final version of the questionnaire made available to users can be found in appendix two.

Designing the questionnaire

Because of the broad scope of the objectives, the purpose of the questionnaire needed to be better defined. The questionnaire therefore consisted of three broad sections: the first dealing with how frequently users employed different information retrieval methods and their role within the community, the second obtaining information as to how important users regarded different types of metadata and the third canvassing users' opinions as to how well different aspects of the information retrieval process worked within bittorrent communities. The questionnaire was required to collect comprehensive data in these areas but also remain short in order to retain respondents' attention – the maxim that 'every question is essential for you to address the research problem' (Leedy and Ormrod 1989) was therefore strictly adhered to.

The majority of questions were closed in order to reduce the time needed to complete the questionnaire and also to simplify coding the data once they had been completed. Several open-ended questions were incorporated to enable additional responses that were not amongst the list of choices present. Many questions make use of summated rating (Likert) scales that allowed respondents to mark a single response on graded series of answers. A five-point scale was used in two instances, both clearly labeled and oriented the same way (negative to positive) to avoid respondent confusion. The second section utilized a series of ten-point scales by which users could assign values to different metadata elements (as identified from the review of bittorrent websites and the test group from the content analysis method).

The website chosen to host the questionnaire was www.questionpro.com. The questionnaire was written using web-based software from this site and given a static web address whereby it could be linked to directly. Completing the questionnaire was undertaken by completing a series of web-forms in html which used javascript to navigate, check and record responses. There were a number of advantages in doing this:

> The questionnaire was presented in a user-friendly and accessible format that did not overwhelm respondents with too much information at one time.

‘Branching’ – where respondents’ answers to questions determined the next question they answered – was handled automatically.

Respondents answers could be validated automatically, ensuring no invalid responses in the final data set.

Answering the questionnaire took less time.

Respondents answering the questionnaire more than once could usually be identified.

Piloting the questionnaire

Once the questionnaire had been written, a pilot was made available online. Six acquaintances were asked to complete it and provide feedback on the issues of usability, clarity and relevance. The pilot group of respondents was chosen on the basis that they had some experience of using bittorrent to distribute material and also, in four cases, experience of carrying out surveys as part of university courses.

Both the results and the feedback returned were useful in refining the questionnaire. Several parts of questions were ambiguous and needed re-wording to clarify their meaning. A major difficulty encountered in the pilot was that information retrieval concepts needed relating more directly to the bittorrent environment. To overcome this, questions were explained in more detail or used examples within the context of bittorrent to better illustrate their point. Although this often resulted in a larger amount of text preceding questions, it was felt this was a necessary compromise of simplicity and layout in order to decrease the difficulty in comprehending questions.

Distribution and sample size

The online questionnaire was updated to the finalised version and then promoted within the bittorrent community. There was no specific sample group targeted for this, the

questionnaire was designed to be answered by any user of bittorrent. In order to try to reach a wide variety of respondents this was done in two ways, through submitting posts on community forums and also through relevant discussion groups that shared information over email. Wherever the questionnaire was promoted, it was made explicit that respondents would remain anonymous, and no identifying information would be asked for.

Two other methods of promoting the questionnaire were also considered and attempted: IRC chat and through the bittorrent website themselves. An initial foray into several IRC chat rooms met with a poor response and was an incredibly time-consuming method of raising the questionnaire's awareness. Many bittorrent sites carry news items on their home page and this was potentially a good means of promoting the questionnaire to the 'casual user' who might not otherwise find posts in website forums or subscribe to mailing lists. Several attempts to contact the administrators of large website were made but unfortunately proved fruitless.

An explanation of the study's purpose and a link to the questionnaire was posted on a total of sixteen bittorrent or file-sharing related forums and two large discussion groups, along with the offer to present the findings of the questionnaire and study once it was completed. The post was always made in a new thread, entitled 'bittorrent research' as posting in existing discussions, while raising visibility, might well cause resentment. An example of the general message posted can be found at the end of appendix two.

It is difficult to estimate how many people viewed the forum posts – in some cases it was met with interest while in two cases the post was censored or deleted by forum moderators. The two discussion groups had approximately 48,000 and 1,000 subscribers although regrettably, there is no way of knowing how many of them are active or read their email. Of this potentially huge sample, a total of 356 people started the questionnaire with 215 people going on to complete it.

Analysis of data

The data was automatically coded by the QuestionPro website software and then downloaded as a comma separated value (csv) file. This was then imported into the SPSS package for further analysis. Frequency tables, averages and variance were calculated for each question and also cross-tabulated in some areas, comparing results based on groups determined by previous answers.