Chapter 2: Background and literature review

2.1 Background

The bittorrent file distribution protocol was developed as an open-source project by Brad Cohen in 2001. It provides an efficient method of distributing data over a peer-to-peer (p2p) network, as an alternative to the traditional server-client model which can be more costly to run in terms of system resources and bandwidth. Bittorrent works through storing metadata in small files (commonly referred to as torrents) which provides the address of a *tracker* or trackers, the file names and sizes being shared and checksums for the data itself. The tracker is a server which maintains a list of the IP addresses of the computers sharing the data and is consulted periodically by the peer computers so they can send and receive data directly between one another.

Users who wish to share data via bittorrent have to create these torrent files. This is a very simple process that requires three things, the address of a working tracker, the data to be shared and software that uses the bittorrent protocol (known as a *client*). Once this is done the user then has to make the torrent file available to others. This can be done in any number of ways but the most common method appears to be to employ websites to index and host these files. The small size of torrent files means that a single website can host many thousands of files simultaneously. The user who uploads the file initially is usually responsible for creating a surrogate record for the torrent file which is linked to the actual file. The format of this surrogate record differs from one site to another – some consist of no more than the filename itself while others have many fields available for users to enter additional information.

2.2 Information-seeking behaviour online

Examining the way in which users locate information is an important part in understanding their interactions with online resources. An often-cited behavioural model of information-seeking is that primarily developed by Ellis. This identified six individual categories of activity: starting, chaining, browsing, differentiating, monitoring and extracting. These were based on the information retrieval methods of a group of social scientists (Ellis 1989). Two similar studies were conducted with other groups of information seekers and identified an

additional four categories: verifying, ending, surveying and filtering (Ellis et al. 1993, Ellis and Haugan 1997).

Ellis" original study was focused on journals and the academic sector. The categories identified have also been successfully applied to the arena of online material, including the world-wide-web. Choo et al. (2000) employ the original six categories and apply „plausible extensions" for use with web documents:

Information Seeking Behaviors and Web Moves						
	Starting	**Chaining**	**Browsing**	**Differentiating**	**Monitoring**	**Extracting**
Literature Search Moves (Ellis et al., 1989, 1993, 1997)	Identifying sources of interest	Following up references found in given material	Scanning tables of contents or headings	Assessing or restricting information according to their usefulness	Receiving regular reports or summaries from selected sources	Systematically working a source to identify material of interest
Anticipated Web Moves	Identifying Web sites/pages containing or pointing to information of interest	Following links on starting pages to other content-related sites	Scanning top-level pages: lists, headings, site maps	Selecting useful pages and sites by bookmarking, printing, copying and pasting, etc.; Choosing differentiated, pre-selected site	Receiving site updates using e.g. push, agents, or profiles; Revisiting 'favorite' sites	Systematically searches a local site to extract information of interest at that site

Adapted from Towards a Behavioral Model of Information Seeking on the Web, Choo et al. (2000)

Fig. 2.1

The information seeking activities or „web moves" here can quite easily be adapted further to deal with bittorrent websites. Ellis" model provides categories of navigational routes that one could apply equally to journal articles, web-pages or torrent sites and records. By using this model when studying an online resource such as bittorrent websites one can determine *if and where* different activities are used by its community and also assess the *potential* for each type of activity to be employed.

These web moves can further be defined by distinguishing between two broad categories: browsing and searching. In reviewing the research available on browsing, Large et al. (2001) find several benefits of browsing over conventional searches – it is a more intuitive form of

activity and also places fewer cognitive demands on its users. Browsing is considered useful when information needs are incomplete or not fully defined. Marchionini (1995, p.49-60) discerns three different degrees of browsing:

> Directed browsing is usually focused towards a specific target and is systematic and often repetitive in its approach (e.g. looking through a departmental website for contact details of one of its members).
>
> Semi-directed or predictive browsing occurs when there is the target is less defined and is less systematic (e.g. the casual examination of a series of records returned by a simple query).
>
> Undirected browsing is where there is no real goal or focus (e.g. skimming through encyclopaedia entries).

While the activity of searching in information retrieval (IR) systems is often characterised as a more methodical and precise technique than browsing, this is not necessarily the case. Casual users often do not make use of advanced search features or strategies, as shown in studies of search engine queries (Spink and Bateman 1998, Spink and Jansen 2004). There also exists some evidence which would indicate that the benefits of using more complex queries in web searches are marginal; Jansen (2000) found there was a very considerable overlap in the results obtained by simple and corresponding complex queries. These studies also showed that users will often employ successive searches in the same subject area. In fact a study of DIALOG users by Spark (1997) was used to form a model of recursive information retrieval by users whereby queries are executed, refined based on feedback received and entered once more. This ongoing process of reformulating queries is an important feature used in any IR system that operates in real-time, including of course the web.

2.3 The role of metadata

Metadata can be broadly defined as „data about data" and fulfils a vital role in an IR system. According to Heery, Powell and Day (1997) metadata can perform any or all of six different functions:

Searching – identifying a resource"s existence.

Location – finding a resource.

Selection – analysing and evaluating a resource.

Semantic interoperability – enabling use of metadata across different domains.

Resource management – managing collections and databases.

Availability – reporting if a resource is available or not.

Although metadata is a term that has found common usage relatively recently, it has a long tradition in the recording and use of bibliographic information. Within libraries, metadata has been used to create surrogate records for items. Safari (2004) rightly surmises that surrogate records have a “limited if almost non-existent, role in the process of web indexing tools” given that the pages are usually indexed as a whole. In the case of bittorrent, however, surrogates are an integral component – due to the size of the material and the wide variety of media types and formats, the entirety of a resource cannot be automatically indexed in full.

Bittorrent resources do share some of the characteristics particular to network resources as identified by Heery (1996); multiple locations, multiple document versions, a lack of stability, the possibility of redundancy, granularity and additional access information. Heery notes that each of these characteristics poses challenges for the effective creation and maintenance of metadata. Particularly relevant to bittorrent resources are the issues of multiple locations, versions and redundancy; p2p networks, and especially the bittorrent protocol, perform more efficiently as more users participate. A multiple number of very similar or identical resources being shared by different groups of users will be detrimental to the overall distribution of the resource.

There has been much research in the area of developing metadata standards which will allow for greater interoperability between domains. The Dublin Core model is one such standard that has garnered considerable interest. Developed initially from the work of Weibel (1997), it consists of fifteen core elements: Title, Creator, Subject, Descriptions, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights (more detailed information can be found online at http://dublincore.org/documents/dces/). These elements are purposely generic in order to encompass a broad spectrum of resources and none of them are mandatory.

In considering the future development of Dublin Core, Lagoze (2001) proposed that efforts to introduce complexity into the model were misguided and counter to its objective of remaining applicable in a wide variety of arenas. Rather than expanding the model to become a monolithic „Esperanto of metadata" he argues that it should be used alongside other more complicated models that are applicable to specific areas. Similarly Duval et al. (2002) acknowledge the difficulties of working with metadata in different spheres:

> "Sources of metadata, like the sources of the resources themselves, will be of different quality and organized around different purposes to reflect the different objectives and business models of information providers."

An all-encompassing metadata standard is then perhaps not achievable or enforceable. Peer-to-peer networks allow any user to distribute any material, regardless of media type or format. Users are required to search through this mass of resources – using the metadata provided by other users – in order to retrieve the desired material. In the case of bittorrent the websites act as intermediaries and user-created metadata can be made (or not) to conform to a standard.

A study by Jansen et al. (2000) compared web search engine queries aimed at finding images, audio or video material with similar studies undertaken previously. The results of the study indicated that multimedia searches placed an additional cognitive load on its users. A „lack of representational congruence" was cited as a likely cause, where both the query and the material indexed were hampered by the difficulty of representing different non-textual media as text. The creators of websites containing multimedia have to decide how to best categorise and describe the material (assuming the media is described at all) in textual metadata while users searching for this material have the task of „second-guessing" the likely

form this metadata will take. For instance, results of the study found that search terms used to identify media type were sometimes in the form of file extensions such as „mpeg", „avi" or „wav".

File extensions or formats are just one instance of additional metadata inherent in digital multimedia resources. In describing digital video Smeaton (2004) cites a long list of metadata present in video media, such as characters, featured objects, scene bounds and colour histograms, stating: “A huge array of metadata can be derived from raw video; this derived metadata permits content-based access to digital video.” Parts of this metadata will be of no help to the casual searcher, but deciding which are relevant and which are not is a difficult task given the subjective nature of user needs and the vast array of different elements that are available in describing non-textual media. Smeaton reports that currently the majority of metadata for video material consists of manual annotation and structural metadata. The focus of recent research however, seems to be on developing and evaluating video surrogates for video material (see for example Wildemuth et al. 2002 and Yang et al. 2003) – thus hopefully eliminating the problem of representational congruence as experienced in textual surrogate records.

2.4 Metadata created by authors and users

The task of creating metadata within formal IR systems has, in the past, fallen to the professionally trained information-specialist. As more information is made available online this becomes increasingly implausible; the ratio of information-specialists to newly generated resources is too small. In addition to this, the dynamic nature of the internet means that resources are subject to frequent changes, as is their corresponding metadata. As a result, there has been some research undertaken recently relating to alternative methods of generating metadata. A small number of author-created metadata records were studied by Greenberg et al. (2001) who found that, by using a simple web form authors are good candidates for creating Dublin Core metadata of an acceptable quality. One aspect not fully explored by this study however, was the possibility of authors misrepresenting their own resource to further their own agenda.

The study conducted by Zhang and Jastram (2006) was interesting in that it identified numerous metadata elements present in web-pages from across four different domains – businesses and industries, library and information science, governmental and non-profit organisations and finally information technology. Overall some elements, such as keyword, description and author, were found to be significantly more popular than others, including date, publisher and resource type. It was also found that there was little use of incorrect or inappropriate metadata on the pages looked at, contrary to the popular concern that authors would misuse metadata in order to promote their popularity. This being said, there *was* evidence that metadata had been created with the intention of becoming visible through search engine queries – there was an emphasis on providing metadata thought to be used by search engines in indexing pages and other metadata elements were sometimes neglected. This was especially the case for commercially orientated sites, where a broad set of keywords would be included in order to increase catchment by search engines. Finally, there was the rather surprising conclusion that the library and information science pages looked at contained a lower quality of metadata than found on other pages.

This last point is perhaps particularly revealing when placed in light of the point made by Robertson (2005) that library and information science workers should be wary of applying traditional metadata practices and standards outside of the traditional LIS environment. He argues that metadata should always be examined within its context and that „metadata quality" cannot therefore be assessed using a single set of criteria. Similarly, Currier et al. (2004) advocates the creation of metadata as a collaborative process between the resource author and a metadata specialist. Through looking at three different case studies of digital learning repositories it was found that resource authors may require specific metadata tools and user support that if not adequately provided for, would decrease the quality of metadata. The trained specialist on the other hand might have difficulty contextualising the information or be impeded by a lack of knowledge about the subject area.

Jones (2005) conducted three case studies on contributor-run archives; The Linux Documentation Project (TLDP), the Degree Confluence Project (DCP) and Etree.org. These were all public, online resources and the studies focused on the collaborative process by which each resource was run. All of them employ different methods of „gate keeping" – the process by which quality is maintained – but rely on volunteers to contribute new material. While TLDP required documents to undergo an editorial process before it would be

accepted, the other two archives were very open with contributions being vetted „post publication" where incorrect records might be then subjected to discussion, editing or removal. Etree itself consists of a number of user-created metadata regarding recordings of concert recordings and the community then uses bittorrent to share the actual recordings between members. Of the 3,941 torrent records created by the community at the time of the study only 88 had been banned with the majority of these being due to a misunderstanding of copyright issues (the site would only allow recordings to be shared where the concert's performers have given permission). The level of rejections by these three projects is reported as being very low – Jones believes that the trust extended to the communities „has been repaid by strong support and adherence to the behaviours defined by the projects". In the case of Etree, he makes the case that peer pressure, pride in one's collection and a willingness to share have led to a generally high standard of quality present in the metadata.

Many bittorrent websites will index the same bittorrent files, although they may have different names and metadata within the records. The way in which different groups of users can – to some extent – determine how a resource is described is similar to the way a „folksonomy" operates. The term folksonomy – derived from folk and taxonomy – has come to mean a system of classification where *users* assign tags or metadata to resources. Many different user tags can be applied to a single resource, meaning that it possesses many access points, each of which can be valid within the context of a particular group of users. Mathes (2004) considers two popular websites that employ user tags in this way, Del.icio.us and Flckr. He describes the limitations of these folksonomies as stemming from its use of uncontrolled vocabulary – causing ambiguity, a lack of uniformity within tags and problems with the use of synonyms. The strengths that Mathes associates with folksonomies are the potential for serendipity and the emergence of „desire lines" (natural paths between resources occurring where community interests overlap). These demonstrate how information retrieval in folksonomies synthesises aspects of conventional searching with browsing. Bittorrent websites and material has the potential to be retrieved in this way, depending on how metadata is organised and used by its community. If users can see who torrents are described by and the language used in doing so, it opens up the possibility for navigation by using shared interests or tastes and community-specific vocabulary.

2.5 Bittorrent and other peer-to-peer networks.

Peer-to-peer (p2p) networks have been in common usage since the development of Napster in 1999. While many similar networks have been created, bittorrent has become one of the more popular according to a recent report conducted by the company Cachelogic (2005). This same report has estimated that p2p networks account for around 60% of all internet traffic, although some criticism has suggested that this statistic is inflated (Sevcik 2005). Another company conducting research into p2p traffic, however, has stated that the number of p2p users has vastly increased since 2003 and, as of November 2005, numbered over 9 million (Mennecke 2005). Giving an accurate assessment of the number of p2p users and the amount of traffic they generate is difficult as there are numerous reports giving conflicting information and pushing different agendas (Ibid.). There does seem to be a general recognition, however, that p2p activity is a significant activity of internet users.

Bittorrent differs from many other p2p networks by the way in which searches are conducted. Other p2p networks have search functionality incorporated into the software, so that queries for material are sent to other peers connected to the network and responses sent back. Material for bittorrent is found by other methods such as on specific websites and then individual p2p networks are set up specifically for this material. A study conducted by Pouwelse et al. (2005) has commented on this hybrid approach to p2p networking which incorporates global components – the tracker and website – as opposed to being completely decentralised. It argues that while system availability may be impeded by the global components, they were, in the case of the website studied, successful in ensuring a high level of quality for content and metadata.[1]

This integrity is related to the existence of „pollution" – material that is not what it claims to be – on the network (Christin et al. 2005). Pouwelse et al. studied a single popular website that indexed and hosted torrent files, as well as running its own tracker, www.suprnova.org. The torrent files listed on this site were maintained by a group of 20 moderators and other

[1] Because much of the material distributed by bittorrent is copyrighted these global components – websites and trackers – are often the targets of legal action. The website studied by Pouwelse et al. has subsequently been taken offline as a consequence of such action, as have several other high-profile bittorrent websites.

volunteers who it was reported managed to „solve the fake file problem". Pollution is thought to be largely the work of commercial organisations, who wish to stop the illegal transfer of commercially produced material (Ibid.) and consists of material that simply does not work. Other malicious content contaminated by a virus or spyware has also been reported (Evers, 2005).

Another issue important to the performance of p2p networks is that of „freeloading" or users who utilise a network to download but do not contribute an equal amount. In a study of the popular Gnutella p2p network, it was found that almost 70% of its users shared no material, while a generous 1% of users provided 50% of all responses to requests for material (Adar and Huberman 2000). The bittorrent protocol goes some way to alleviating this problem by attempting to keep users download speed in proportion to their upload speed (Cohen 2003). The problem still remains however as people may cease downloading when they have a complete copy of the material, well before they have uploaded an equal amount of material to that downloaded. The problem is made worse by the fact that internet connections often do not have upload capabilities equal to downloading speeds (Cachelogic 2005) meaning that achieving a „share ratio" (amount downloaded to amount uploaded) of 1:1 can take a longer time.

Hales and Patarin (2005) hypothesise that the problem of free-riding over bittorrent networks is overcome by the activity of „group selection" undertaken by its users. Because material is available through multiple sources (different torrent files and trackers for the same material), users will naturally move to the sources that provide better performance. Thus sources containing selfish users will dwindle and sources that have altruistic users will flourish. This model can be extended to communities of users based around bittorrent websites – those that provide good distribution of torrents will attract a greater number of users and vice versa.

2.6 „Warez" users and communities:

The bittorrent protocol can be used to share any material but has attracted the most of its attention in the media from the fact that it is used to share pirate material (where the user

does not legally have the right to use the material because of its copyright status). While the legality of such sharing falls beyond the scope of this project, it is prudent to further understand a little of such users" and communities" activities, given that this material arguably makes up the majority of bittorrent traffic. It is in software piracy that piracy culture has its roots; the term „warez" itself was initially used to refer to illegal copies of computer software, although now can be attributed to any form of digitally pirated material. The website www.defacto2.net currently maintains an archive of documents produced by groups and individuals within this culture and extends as far back as 1986 (at this time, „online" piracy was limited to using bulletin board systems to exchange material). The warez scene, sometimes just referred to as „the scene" has developed certain characteristics, some of which are relevant to users of p2p networks.

A very comprehensive description of this community is given by Rehn (2001) in which he details the groups and vocabulary used by its members. Most importantly, he describes the groups of users known as „release groups". These are the suppliers of warez; "an organisation that obtains programs, manipulates them and releases them in packaged form" (p. 32). When material is released by these groups, an "NFO" file (an ASCII text file with the extension .NFO) is also released identifying the release group responsible. This file may also contain additional information, such as the group"s members and affiliations as well as details specific to the software. Warez material may not be released directly onto p2p networks; according to one account there is an identifiable „food chain" whereby material is shared by several different types of users until it filters down to p2p networks, located at the bottom („2600", 2004).

Of the groups identified amongst the warez scene, release groups are the most directly relevant to this study as they are often listed within torrent records of pirated material. Other groups are largely concerned with distributing the material between interested parties. In a study of online music piracy, Cooper and Harrison (2001) found that there existed a social hierarchy within those trading music, based on a number of factors including bandwidth, storage capacity and access to new („zero-day") releases. This study, however, was conducted on piracy that took place on Internet Relay Chat (IRC) channels. This method of online piracy, although still popular today, is very different from bittorrent and other p2p networks in that it usually requires a greater degree of social interaction. Users are able to directly speak to one another and allocate bandwidth and resources accordingly.

Conversely, bittorrent is customisable to a much lesser degree and the sharing of information is largely automated, requiring a much lower social investment from its users.

2.7 Conclusion

Peer-to-peer networks are now an established internet technology and form a substantial part of the activity conducted over the internet. Bittorrent is one of the more recent and popular iterations of this technology that increases the potential for media to be distributed amongst internet users more efficiently and speedily. Unlike many other p2p networks it requires its users to provide a means of locating material to be shared and also providing new material. Bittorrent websites hope to provide this service which involves two primary components: providing a framework for internet users to house and locate torrent material and enlisting the community to populate this framework with torrent records.

This framework has to provide a reliable method for its users to find appropriate records and as such can be compared to other online IR systems in its features, performance and the behaviour of its users. At the centre of this framework there is the metadata of the torrent records themselves, often provided by the users. While there have been several studies of author and user generated metadata these have usually been within different settings and dealt with textual media. A study conducted of bittorrent websites needs to focus on the user, keeping the IR functionality and performance of the system itself within the social context of the users, who essentially drive it and make it work. In studying bittorrent one can observe communities utilising new and existing systems of information organisation in order to meet their needs and best harness the potential of a new technology.