

VideoTag: encouraging the effective tagging of internet videos through tagging games

Item Type	Thesis or dissertation
Authors	Lewis, Stacey.
Citation	Lewis, S. (2014) VideoTag: encouraging the effective tagging of internet videos through tagging games. University of Wolverhampton. http://hdl.handle.net/2436/621745
Download date	2026-05-17 02:35:08
Link to Item	http://hdl.handle.net/2436/621745

VideoTag: Encouraging the Effective Tagging of Internet Videos Through Tagging Games

Stacey Lewis BA (hons) MSc

A thesis submitted in partial fulfillment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy

May 2014

Director of Studies – Professor Mike Thelwall

Supervisor - Dr Kevan Buckley

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgments, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

The right of Stacey Lewis to be identified as author of this work is asserted in accordance with ss.77 and 78 of the Copyright, Designs and Patents Act 1988. At this date copyright is owned by the author.

Abstract

The tags and descriptions entered by video owners in video sharing sites are typically inadequate for retrieval purposes, yet the majority of video search still uses this text. This problem is escalating due to the ease with which users can self-publish videos, generating masses that are poorly labelled and poorly described. This thesis investigates how users tag videos and whether video tagging games can solve this problem by generating useful sets of tags.

A preliminary study investigated tags in two social video sharing sites, YouTube and Viddler. YouTube contained many irrelevant tags because the system does not encourage users to tag their videos and does not promote tags as useful. In contrast, using tags as the sole means of categorisation in Viddler motivated users to enter a higher proportion of relevant tags. Poor tags were found in both systems, however, highlighting the need to improve video tagging.

In order to give users incentives to tag videos, the VideoTag project in this thesis developed two tagging games, Golden Tag and Top Tag, and one non-game tagging system, Simply Tag, and conducted two experiments with them. In the first experiment VideoTag was a portal to play video tagging games whereas in the second experiment it was a portal to curate collections of special interest videos. Users preferred to tag videos using games, generating tags that were relevant to the videos and that covered a range of tag types that were descriptive of the video content at a predominately specific, objective level. Users were motivated by interest in the content rather than by game elements, and content had an effect on the tag types used. In each experiment, users predominately tagged videos using objective language, with a tendency to use specific rather than basic tags. There was a significant difference between the types of tags entered in the games and in Simply Tag, with more basic, objective vocabulary entered into the games and more specific, objective language entered into the non-game system. Subjective tags were rare but were more frequent in Simply Tag. Gameplay also had an influence on the types of tags entered; Top Tag generated more basic tags and Golden Tag generated more specific and subjective tags.

Users were not attracted to use VideoTag by the games alone. Game mechanics had little impact on motivations to use the system. VideoTag used YouTube videos, but could not upload the tags to YouTube and so users could see no benefit for the tags they entered, reducing participation. Specific interest content was more of a motivator for use than games or tagging and that this warrants further research. In the current game-saturated climate, gamification of a video tagging system may therefore be most successful for collections of videos that already have a committed user base.

Table of Contents

1	Introduction	1
1.1	Thesis Structure	7
2	Literature Review	10
2.1	Introduction	10
2.2	Online Video	10
2.2.1	YouTube Studies	11
2.2.2	Video Search	16
2.2.3	Video Annotation – Bridging the Semantic Gap	18
2.2.4	How Users Search for Videos	22
2.3	Games With A Purpose	25
2.3.1	Video Tagging Games	34
2.3.2	Crowdsourcing and Motivation to Participate	38
2.3.3	Theory of Play - Motivation, Engagement and Flow	44
2.4	Defining ‘Casual’ in Game Design	53
2.4.1	Designing for Player Enjoyment – Player Type	56
2.4.2	Designing for Player Enjoyment - Fun Factors	59
2.4.3	Measuring Enjoyment	63
2.5	Gamification	66
2.6	Tagging	73
2.6.1	Types of Tag	81
2.6.2	Motivation to Tag	85
2.6.3	Tagging Video	88
2.7	Summary	92
3	Research Questions	94
4	Classification Studies – Straight Tagging of User Generated Video	98

4.1	Tagging YouTube	98
4.1.1	Introduction.....	98
4.1.2	Literature Review	98
4.1.3	Research questions	100
4.1.4	Methods	101
4.1.4.1	Data Collection	101
4.1.4.2	Classification Scheme.....	101
4.1.4.3	Findings.....	105
4.1.5	Discussion.....	109
4.1.6	Conclusion	110
4.2	Broad Video Tagging.....	112
4.2.1	Introduction.....	112
4.2.2	Methods	114
4.2.3	Data Collection.....	114
4.2.4	Findings	116
4.2.4.1	Comparison of Tagging Systems	116
4.2.4.2	Tag Classification.	120
4.3	Conclusion	127
5	The VideoTag Experiment	129
5.1	Introduction	129
5.2	Primary Design	133
5.2.1	The VideoTag Website Design.....	133
5.2.2	Golden Tag.....	137
5.2.2.1	Synopsis	137
5.2.3	Top Tag.....	142
5.2.3.1	Synopsis	142
5.2.4	Simply Tag.....	146
5.3	Methods	151
5.4	The Phase One Experiment: Publicity to attract users.....	153
5.4.1	Methods	153
5.4.2	Results	154

5.4.2.1	Usage statistics	154
5.4.2.2	Tag Frequency.....	155
5.4.2.3	Observations.....	158
5.4.3	Discussion.....	159
5.5	Phase Two Design	161
5.6	Phase Two Prototype – SciFest Experiment	170
5.6.1	Methods	170
5.6.2	Results	170
5.6.3	Design Decisions Based On Analysis of the SciFest Prototype	175
5.6.4	Summary.....	178
5.7	Phase Two Experiment	179
5.7.1	Methods	179
5.7.2	Results	180
5.7.2.1	Usage Statistics	180
5.7.2.2	Tag Frequency.....	181
5.7.3	Discussion.....	185
5.8	System Comparisons	186
5.8.1	Games vs. Non-Game.....	186
5.8.2	Phase one Experiment vs. Phase two Experiment	191
5.8.3	Motivation.....	193
5.9	Conclusions.....	195
6	Usability and Engagement Evaluation 198	
6.1	Introduction	198
6.2	Usability Evaluation	206
6.2.1	Methods	206
6.2.2	Results	209
6.3	User Satisfaction	213
6.3.1	Methods	213
6.3.2	Results	214
6.4	Playability.....	221
6.4.1	Methods	221

6.4.2	Results	222
6.5	Engagement and Enjoyment.....	229
6.5.1	Methods	232
6.5.2	Results	233
6.5.3	Discussion.....	236
6.6	Barriers to Use or Play	241
7	Classification Studies – Game Based Tagging of User Generated Video	244
7.1	Introduction	244
7.2	Methods	250
7.3	The Classification Scheme	250
7.4	Reliability of the Tag Classification Scheme	255
7.5	Selecting Tags for Classification	255
7.6	Inter-coder Test for Reliability	258
7.7	Results.....	258
7.7.1	General Observations	258
7.7.2	Phase One and Phase Two	260
7.7.3	Golden Tag vs. Top Tag	269
7.7.4	Game vs. Non-Game (Simply Tag).....	273
7.7.5	Entertainment vs. Informative.....	276
7.8	Discussion	279
7.9	Conclusion	282
8	Conclusion	287
8.1	Introduction	287
8.2	Research Objectives	288
8.2.1	General Objectives	288
8.2.2	Specific Objectives.....	289
8.2.2.1	R1 - Which game elements that make casual games engaging also help to make video tagging games engaging?.....	290

8.2.2.2	R2 - Can game elements affect the types of tag that players enter?	295
8.2.2.3	R3 - Does video content affect tag type?	298
8.2.2.4	R4 - Can video tagging games encourage users to enter specific level descriptive tags as well as general level descriptive tags?	301
8.3	Conclusions and Future Work	303
8.4	Contribution to Knowledge	304
9	References	306
Appendix A 336		
	Research in progress paper published and presented at ISSI 2009.....	336
Appendix B 346		
	Poster presented at SNIC 2009.....	346
Appendix C 348		
	Usability Questionnaire SUS evaluation.....	348
	Desirability Questionnaire	349
	Playability Questionnaire.....	350
Appendix D 351		
	VideoTag Tag Classification Instructions	351

List of Tables

Table 2-1 Mapping fun factors to player types.....	63
Table 2-2 Summary of tagging system characteristics.....	75
Table 2-3 The vocabulary problem for tagging systems.	78
2-4 Example tags categorised by basic level theory and Panofsky's levels of meaning.	83
2-5 Summary of concepts describing motivations to tag.	87
Table 4-1 The tag classification scheme including category definitions - adapted from (Angus et al. 2008).	103
Table 4-2 Total number of tags and corresponding percentage of all tags.....	106
Table 4-3 Category Groupings.	114
Table 4-4 The breakdown of videos over the eight categories.....	115
Table 4-5 The percentage of tags assigned to Entertainment and Informative videos in YouTube and Viddler.....	121
Table 4-6 The amount of tags in each tag type classification category for YouTube and Viddler.....	124
Table 4-7 The number of tags in each tag type classification category for Entertainment and Informative videos.....	125
Table 5-1 Web traffic statistics for the phase one experiment.	155
Table 5-2 The ten tags with highest user agreement for individual videos in phase one.....	157
Table 5-3 The amount of videos tagged in each category.	157
Table 5-4 The number of videos in each phase two category.	163
Table 5-5 The mean, mode and maximum tags entered in each game or Simply Tag session.	171
Table 5-6 The number of tags assigned to each of the 14 videos used in SciFest.	172
Table 5-7 Ten most entered tags in each system during SciFest.....	173
Table 5-8 Web traffic during the phase two experiment (N.B., June also includes phase one and SciFest traffic).....	181
Table 5-9 Number of tags entered using each system during the phase two experiment.....	182

Table 5-10 The most tagged categories in phase two.....	183
Table 5-11 The ten most frequently entered tags in each system phase two.....	183
Table 5-12 A comparison of the mean, mode and maximum number of tags entered per video during phase one and phase two.	184
Table 5-13 Comparison of the mean, mode and maximum number of tags entered per game in each system.....	188
Table 5-14 Ten most frequently entered tags in Golden Tag, Top Tag and Simply Tag.....	189
Table 5-15 The number of high frequency tags entered in each experiment.	192
Table 5-16 User motivation to use VideoTag	194
Table 6-1 Results of the Cronbach’s alpha test for reliability for the SUS questionnaire.	212
Table 6-2 Frequency of words entered in the MPRC evaluation.	215
Table 6-3 The mean, mode and median of responses to the playability questionnaire (n=7).	224
Table 6-4 Grouping structure of questions and summary of findings.....	225
Table 6-5 Pearson correlations for each question pair (n=7).....	226
Table 6-6 Paired t-tests for differences in mean between question categories (significant results are highlighted in grey) (n=7).	228
Table 6-7 Results of the four enjoyment evaluations using the Game Flow model (Sweetser and Wyeth, 2005).	233
Table 7-1 The classification scheme detailing each tag type category.....	251
Table 7-2 The amount of tags classified in each condition.	257
Table 7-3 Tag type categorised by Objective and Subjective Vocabulary.....	258
Table 7-4 Percentages of tag types in the phase one and phase two experiments.....	261
Table 7-5 Results of z-test for differences in proportion tests for each tag type category between phase one and phase two.	264
Table 7-6 Results of z-test for differences in proportion tests for each tag type category between phase one and phase two with SciFest data removed.....	265

Table 7-7 Significant results from difference of proportion z-tests comparing basic and specific tag type classifications from phase two including SciFest data and phase two excluding SciFest data.....	268
Table 7-8 Results of z-test for differences in proportion tests for each tag type category entered using Golden Tag and Top Tag.	271
Table 7-9 Results of z-test for differences in proportion tests for each tag type category entered using either a game or Simply Tag.	275
Table 7-10 Results of z-test for differences in proportion tests for each tag type category entered for Entertainment and Informative videos.	277

List of Figures

Figure 4-1 - Frequency distribution of tags per video on Viddler (log-log scale).....	117
Figure 4-2 - Frequency distribution of tags per video on YouTube (log-log scale).....	118
Figure 4-3 Comparison of tags and views on YouTube and Viddler.....	119
Figure 4-4 A comparison between YouTube and Viddler tags.	120
Figure 4-5 The proportions of tag type assigned to entertainment videos.	122
Figure 4-6 The proportions of tag type assigned to informative videos.....	123
Figure 5-1 The phase one VideoTag homepage design.	135
Figure 5-2 The Golden Tag in game interface.	141
Figure 5-3 The Golden Tag end game interface.....	141
Figure 5-4 The Top Tag in game interface.....	144
Figure 5-5 The Top Tag interface showing success during a game.	144
Figure 5-6 The Top Tag select a category and game-over screens.	146
Figure 5-7 The Simply Tag interface to browse by tag.....	149
Figure 5-8 The Simply Tag interface to browse for videos.....	149
Figure 5-9 The Simply Tag interface to tag a video.....	150
Figure 5-10 The Simply Tag end of session interface.....	151
Figure 5-11 Rank tag frequency in phase one (log-log scale).....	156
Figure 5-12 Rank frequency of the amount of tags per video in phase one.	158
Figure 5-13 The homepage of the phase two version of VideoTag.	167
Figure 5-14 The select a video page in phase two.....	168
Figure 5-15 The phase two game over page.....	168
Figure 5-16 Phase two changes to Simply Tag, navigation buttons replace the select video panel.....	169
Figure 5-17 Phase two changes to the finish tagging page in Simply Tag.....	169
Figure 5-18 Amount of tags per video, per tagging method, in Crazy Science Experiments category (see Table 5-6 for the titles of each of the 14 videos).....	171

Figure 5-19 Rank frequency of tags entered into the three tagging systems during SciFest.	175
Figure 5-20 Selecting a tagging system in phase two.	177
Figure 5-21 Rank frequency of the amount of tags per video.	185
Figure 5-22 Player Retention Curve – rank frequency of the number of games players played.	187
Figure 5-23 Total tag frequency in the three tagging systems.	191
Figure 6-1 A box plot of SUS scores.....	213
Figure 6-2 Frequency Distribution of MPRC words.....	217
Figure 6-3 Frequency of positive and negative MPRC words.	218
Figure 6-4 Frequency of enjoyment and usability MPRC words.....	219
Figure 6-5 Amount of positive and negative MPRC words grouped by enjoyment and usability.	220
Figure 7-1 Overall distribution of subjective and objective tags.....	259
Figure 7-2 Overall distribution of specific objective and basic objective tags	260
Figure 7-3 Distribution of tag type in Golden Tag and Top Tag	269
Figure 7-4 Distribution of tag type in Game (Golden Tag and Top Tag) and Non-Game (Simply Tag) systems	274

1 Introduction

The last ten years have seen a shift in the way that people use the internet and how content is generated. There has been a movement away from the read-only web that was the realm of computer programmers, to a social web where the users of a website produce the content, “The passive user is now an active producer of content” Silva and Dix (2007). Although individuals once had to rely on web designers and developers to produce content, now there is a multitude of websites offering users the opportunity to self-publish. Examples include blog sites, review sites, Facebook, Twitter, Flickr, Instagram, YouTube and even sites that let you create your own free website. This has seen a rise in individuals sharing content but has also changed how companies market their products, with social media being a new advertising platform. There are now many micro blogs, status updates, photos, videos, blogs and sound bites for web users instead of a collection of homepages.

Not only has the way that content is created changed, the way that users consume content has also changed. The rise in portable handheld devices coupled with improved wireless and mobile internet networks means that users are no longer tied to a desktop PC or laptop. Smart Phones and Tablets equipped with good quality cameras have created a new way for people to communicate and publish themselves online. Users can share photos and videos in seconds on websites such as Facebook, Twitter, Instagram, Vine and YouTube. YouTube has had phenomenal success since its launch in 2005, being owned by Google since 2006. Over 6 billion hours of video are watched each month by more than one billion unique visitors, with 100 hours of video being uploaded to YouTube every minute (YouTube, 2013). The year 2013 saw a surge in mobile video sharing with the launch in January of Twitter’s video sharing app Vine (acquired by Twitter in October 2012) and the subsequent introduction of video to Facebook’s photo sharing app Instagram in June of the same year. Vine

allows users to upload and share short 6 second video clips, whereas Instagram allows 15 second clips. Currently over 50% of mobile web traffic is video and it is predicted this will increase to two-thirds by 2017 (Neomobile, 2013). 71% of internet users in the USA watch videos online (Rainie, 2014); in the UK 86% of internet users visit a video site at least once a month, watching 240 million hours of video. YouTube receives 70% of that traffic (Experian, 2011). It is the largest video sharing site and the third most visited website globally (Alexa, 2013). It is not clear if the increasing popularity of services like Instagram and Vine will reduce YouTube's video traffic share.

Users can archive their own videos and share videos within social networks but typically with limited textual data associated, with each one there is a need to index all this user-generated video to make it easier for other users to find. Users who upload video to sharing sites want it to be seen, and views and likes have replaced hits as "badges of honour" for today's web user (Groh, 2012). There are presumably millions of unwatched videos on YouTube, Instagram and Vine. YouTube search is currently text based; the majority of online video can only be queried based on the textual information associated with it. When searching for a video a user can only search textual representations of objects or textual representations of opinions or interpretations of the video content. This is limited as there is typically not enough textual data available to fully describe YouTube videos (Trant, 2009). No major search engine offers content based search for video as it is currently not accurate enough or adaptable enough to large scale corpora (Müller et al., 2012). There are few studies that investigate the effectiveness of YouTube search and the majority of research concentrates on generic search engines or small custom video libraries. Nevertheless current methods of video search are insufficient because textual

descriptions of video content are inadequate. This will be discussed further in Section 2.2.2.

Digital video libraries in contrast to social web video sites, are likely to be catalogued and indexed by accurate descriptions, albeit potentially not detailed enough (Trant, 2009). User generated video is typically uploaded by inexperienced classifiers, with poor descriptions and poor titles, if titled or described at all and yet videos that are largely returned by a YouTube search tend to be the videos that have the most textual descriptions (Halvey and Keane, 2007). These textual descriptions are mostly inaccurate Kern *et al.* (2008), perhaps added to improve the chances of being found and watched regardless of relevance to content. Professional classifiers cannot be employed to classify the millions of user generated YouTube videos. Automatic methods of video annotation offer a solution although current methods are not sophisticated enough (Morsillo et al., 2010). Only low level features at a perceptual level (e.g., colour, shape and texture) can currently be extracted from the video content. What are required are high level semantic descriptions at a conceptual and abstract level that accurately describes the content. Only humans are currently capable of extracting high level semantic descriptions from videos. To employ annotators is too expensive and financial incentives may affect the quality of their input if they are not enthusiastic about the activity (Mason and Watts, 2010). A potential solution is to encourage users to annotate the videos that they watch and upload.

Considerable research has been conducted on image tags, but little research has been done on tagging videos. The types of tag, language classification, use for search,

improving search and motivations to tag have all been studied. Image search has made more progress than has video search, aided by the work of Von Ahn and Dabbish (2004) and Google's implementation of their ESP Game as Google Image Labeller. Just as techniques for computer image interpretation are being applied to research for video content search, techniques used in evaluating image tags could be applied to video tags. The majority of image studies have been conducted on Flickr which during the 2000s in particular was a popular photo sharing site that used tagging as a way of categorising and organising collections of user generated images. The tags were public and available via an API so it was easy to investigate. YouTube and Viddler (until 2014) are the closest comparable video sharing sites, although both are less sophisticated in terms of tagging. This thesis presents two preliminary studies on tagging behaviour and the types of tags in these systems (See Chapter 4).

Tags can hold different types of descriptions for video, from a basic label like 'woman' for instance, to labelling the woman as 'Beyonce' and expressing a positive or negative opinion. All of the textual data generated through the tagging of a video can be useful in a number of applications: classification, categorisation, descriptions for visually impaired, indexing and search (Gligorov et al., 2013). Tagging, whilst still existing in its original format, has been popularised into a new web phenomena: #hashtags. The first use of the hashtag on Twitter is reported to be 2007 (MacArthur, 2014). They are entered not to order tweets for future retrieval but as a conversational tool, grouping tweets into conversations or free form topics, adding emphasis, or improving visibility. Guy and Tonkin (2006) found evidence of the # symbol preceding tags in del.icio.us and surmised that their use was to improve visibility of the tag. Huang *et al.* (2010) state that hashtags are conversational rather than organisational, but whilst the organisation might not be for later retrieval it is present

to some degree in filtering and directing content. The hashtag has spread from Twitter into Instagram and is currently emerging on YouTube and Facebook; it has also had an impact on colloquial speech. Organisations, in particular media organisations, use hashtags to aggregate user generated content; they can collect comments, photos and videos from various social networks using the same hashtag. This phenomenon has brought tagging into the mainstream, but there is no evidence to suggest that the majority of users have an understanding of its benefits for information retrieval. There is little academic research on hashtagging in any network other than Twitter. The current popularity of the hashtag affords validation to this research by indicating that tagging is evolving, web users still use it and are open to using it, albeit in an evolved form. Nevertheless “To reap the benefits of tagging people need to be persuaded to tag the resources they consume” Melenhorst and van Velsen (2010). If tagging is the solution to the video retrieval problem, how can users be encouraged to tag what they watch and upload and to tag effectively?

Bouca (2012) describes the ‘ludification’ of culture, a society where play is a centric element. We are less work focussed and more play focussed, we have more leisure time and more emphasis is placed upon it. Games are infiltrating all aspects of life, with people becoming more accustomed to reward structures in daily routines (McGonigal, 2011). This process is referred to as gamification (described in Section 2.5). Most game experience models, player type models and guidelines for designing games concentrate on games as a whole. There is little focus on the differences between console games, (e.g., Grand Theft Auto), MMORGs (e.g., World of Warcraft), downloadable PC games (e.g., Bejewelled or Peggle), mobile games (e.g., Candy Crush Saga or Temple Run), or non-computer games (e.g., Monopoly, Basketball or Chess). Game mechanics and player types are identifiable in all these games. Player

behaviour can be modelled independent of the type of game; players will have different objectives and motivations for the type of game they pick and have an experience in mind. The role of the game designer is to apply game mechanics that meet these player requirements. The most successful games will meet many requirements in one form or other. Video tagging is not an obvious choice for a game but users are willing to devote time to tagging if the process is made more enjoyable in a game format (Von Ahn and Dabbish, 2004). The Games With A Purpose (GWAP) concept adapted from serious games research has been developed extensively to make mundane activities more engaging (discussed in more detail in Section 2.3). In a spectrum with serious games at one extreme and gamification at the other, GWAP would be positioned somewhere between the two opposing extremes. Whilst GWAP are not serious games, they have more characteristics of a standalone game than a gamified system with a layer of game elements. The current problem for GWAP is that the systems are no longer novel. The ESP Game was perhaps successful because of its novelty but the ESP Game model has been replicated extensively and is no longer novel enough to attract players on the same scale. A current challenge is to improve the model by taking methods from casual game design to attract players.

This research focuses on generating high quality textual data for YouTube videos by employing the GWAP concept to motivate users to tag videos. It investigates what elements of casual game design can be applied to the video tagging interface to encourage users to tag. User motivation to engage in a video tagging game is considered from different angles; their motivation to tag, their motivation to watch videos and their motivation to play games. Through a literature review, game mechanics will be identified that could be applied to the process of tagging a video to potentially optimise enjoyment for users and to promote and maintain use. The

research will evaluate whether gamification can be successful when applied to video tagging, although cognitive cost of watching a video, tagging and playing a game may be too high for a game to be engaging. Many tagging game projects measure success based on user numbers and the quantity of tags, rather than investigating the quality of output the games generate. Tagging games need to focus on tag diversity not just tag quantity (Melenhorst and van Velsen, 2010; Chi and Mytkowicz, 2008). Throughout this research consideration is given to the types of tag that users assign to videos, focussing on how accurately they describe the video content and the language used. Through the literature review a definition will be formed of what constitutes tag quality in relation to improving textual data. The literature review will inform the design of research experiments to explore which elements of gameplay can encourage users to tag effectively so that the descriptions entered can be used to improve textual data for videos.

1.1 Thesis Structure

The general objective of this research is to develop video tagging games to investigate what aspects of gameplay will encourage users to tag videos and to tag them effectively. A set of tags are effective in this sense if they describe the content of the video with a range of words that can be used to significantly improve video search. Since the area of video tagging games is new, the scope of the literature review is wide and multidisciplinary. Literature will be gathered from the fields of: YouTube studies, video retrieval, casual game design, theory of play, games with a purpose, gamification, motivation theory, tagging systems, tagging of visual resources and tag classification (primarily image tagging). The literature review, presented in Chapter 2 will apply general themes from these research areas to the

specific context of video tagging games. The findings will help formulate the specific research questions as outlined in Chapter 3.

The outcome of preliminary studies into tagging practice on video sharing websites will be reported in Chapter 4. Two classification studies for tags entered on YouTube and Viddler will be evaluated. The studies will provide an understanding of how users tag videos in video sharing systems and the types of tag they enter. The findings will inform the requirements for the design of the main research experiment, revealing the types of tag that VideoTag should encourage. Chapter 5 details the design and implementation process for the VideoTag project, outlining aspects of its design that were informed by the literature review and preliminary studies. These include which elements of gameplay will attract use and encourage tags of a certain type; strategies to motivate use and which player types certain game elements will attract. This chapter contains the primary methods and results for the VideoTag project. Design decisions for each system and experiment are described, explaining how the results from one experiment informed the design of the other. Rudimentary results are discussed concentrating on usage statistics and tag frequency analysis. The results presented in Chapter 5 form the basis for further investigation of the data. Chapter 6 presents results and a discussion of three user studies (one usability, one playability and one user experience questionnaire) conducted during the second experiment as well as the findings of developer evaluations of the systems used in each experiment to assess game enjoyment. Discussion centres on user motivation and aspects of the system that encouraged or deterred use. Chapter 7 presents the findings of tag classification studies conducted on tags generated throughout the project. The classification scheme used in the preliminary studies will be used for cross evaluations with the preliminary study results. Various testing conditions will

be created based on the findings of the literature review and the specific research questions, as well as trends highlighted from the tag frequency analysis in Chapter 5. Finally, Chapter 8 concludes the thesis.

2 Literature Review

2.1 Introduction

The general introduction identified three broad research areas that provide problems for the specific area of Video Tagging Games; Video Search, Games With A Purpose (GWAP) and Tagging. These will be discussed in more detail in this chapter. The literature review also provides background for design decisions and research methods discussed in subsequent chapters. Section 1 provides context for the problems for Video Search, discussing relevant research in YouTube studies, current methods of video search, current thinking on bridging the semantic gap and how users search for video. Section 2 describes GWAP in detail, reviewing relevant GWAP projects and detailing existing video tagging games. This section expands the specific research area to identify what motivates users to participate in GWAP's. Section 2 concludes with an overview of the theory of play, identifying what play is, why people play and how they engage with games; where possible relating theory to GWAP. Section 3 examines casual game design theory, highlighting types of player, how to design for player enjoyment and how to measure enjoyment. Section 4 investigates gamification and how this emerging area is related to and incorporates GWAP research. Finally Section 5 explores tagging research, focusing on tagging system design, the types of tag users enter and motivation to tag. It also reviews the limited research in tagging video.

2.2 Online Video

An abundance of online video, both user generated and professionally produced, is continuing to grow exponentially. There is a wealth of research available on the problems of indexing video, categorising collections, user behaviour, advertising,

machine reading and automatic and manual annotation. The growth of online video and the lack of sufficient solutions to the problems of video search keeps these research areas active. This section provides an overview of academic studies on YouTube, how users search for video, key research areas in improving video search and problems still to be solved.

2.2.1 YouTube Studies

YouTube may be branded as a sharing site for user generated content but it is undoubtedly the perfect channel for advertising. In order to prevent continuous lawsuits for breaching copyright, YouTube has had to partner with mainstream media organisations. Equally, in order to try and combat copyright infringement, mainstream media has had to embrace YouTube, (Andrejevic, 2009). Despite huge interest in monetizing YouTube, only 3% of all YouTube clips contain advertising (Kim, 2012). Users can buy a place in the featured video section of the homepage, they can buy keywords, and adverts can be placed at the beginning of videos. YouTube also uses Google AdSense¹ to show adverts to users based on their Google searches, but this has only managed to monetize 14% of views (Ulges et al., 2013). Advertisers do not want their brand associated with low quality user generated video and viewers could take a lack of adverts as an indication that the video is not worth watching (Kim, 2012). User generated video without adverts is not as highly ranked or publicised by YouTube metrics (Kim, 2012), which creates a gap between User Generated Content (UGC) and Professional Generated Content (PGC) (Andrejevic, 2009). Being content uploaded by mainstream media organisations, PGC

¹ www.google.com/adsense/

is generally advert heavy and may be region restricted e.g., can only be played if in the US. Kim (2012) notes how early visitors to YouTube would recognise the increasing dominance of PGC. There is little research into whether this increase in PGC and advertising affects user experience. Burgess and Green (2009) found that in a sample of 4320 most popular videos retrieved over 3 months in 2007, only one third of the 'most viewed' videos were user generated. In contrast, they found that for ranking categories where social engagement was quantified over views the reverse was true, with two thirds of the videos being user generated. 42% of video content was categorised as being PGC although, when looking at the uploader of the video, 61% were by individuals or 'users', 20% by small enterprises or independent firms and only 8% by big companies. However, in the six years since this study it is plausible that these ratios could have changed. Cost of video hosting and pressure from mainstream media to make their videos more prominent and to reduce copyright infringement has forced YouTube to invest in developing interfaces, tools and features that have made PGC content more dominant (Kim, 2012).

Mainstream media view YouTube as another platform for distributing their content; however, as YouTube's structure is social, views are reliant on engagement with users (Shamma et al., 2007). Advertising needs to be targeted at specific users or specific content. Without good, useful descriptions of the content or information about the users watching the content, it is difficult to effectively monetize YouTube. Ulges *et al.* (2013) highlight how an improvement in the textual data associated with a video could improve user experience by making it easier to assign adverts to videos that contain similar content. There would then be a higher probability of the advert being relevant to the user. In particular, if textual descriptions for UGC were improved advertisers may find more videos that suit their brands, therefore allowing

more UGC to be promoted on YouTube. Halvey and Keane (2007) found that pages containing promoted videos had more textual data than regular pages. They argue that videos are promoted if they have a large number of views, rather than having a large number of views because they were promoted. This would suggest that users of YouTube will not watch a video just because mainstream media, a brand or an individual has paid YouTube to promote it; they watch a video because other users have watched it. Views are votes of recommendation. Ranking videos based on view count assumes all users want to watch popular content. Figueiredo *et al.* (2011) and Tao *et al.* (2012) note that the most viewed videos are not necessarily the best videos, meaning that millions of videos are going unwatched or are only relevant to a few users. How do those few users find videos of more specific interest? Is it possible to improve access to the majority of content that is rarely viewed?

Cha *et al.* (2007) and Capra *et al.* (2008) highlight the potential for improving the findability of the unlimited number of non-popular videos and videos of specific interest. Marchionini *et al.* (2009) found that niche content videos had the fewest views. Li *et al.* (2012) highlight the difficulty in finding niche content because of a lack of descriptive text and poor categorisation. Li *et al.* (2012), Cha *et al.* (2007), Marchionini *et al.* (2009) and Paolillo and Penumarthy (2007) discuss methods to find niche videos and videos of specific rather than generalised interest on YouTube and the problems associated with this. The authors agree that a video would have a higher number of views when the uploader utilised the social elements of YouTube for promotion or when users watching the video engaged in these social features. However, the authors found that few users interact with these social features. Social features of YouTube include uploading videos, sharing of videos on other social networks, e.g., Facebook or Twitter, embedding videos in blogs or web pages, rating

videos, commenting on videos and tagging videos (entering keywords that describe the content). Each of these features provides additional textual content that can be useful in improving the accuracy of video search. Both Halvey and Keane (2007) and Cheng *et al.* (2008) found that a user's main motivation to visit YouTube is to watch videos and that few users engage with the social features of the site. They also found that the majority of videos are uploaded by only a few users. Many authors discuss evidence of the Pareto Principle (Newman, 2005) in user activity on YouTube. Gill *et al.* (2007) discovered that 20% of the most popular videos are watched by 80% of YouTube users. Similarly, Cha *et al.* (2007) observed that 10% of the most popular videos in their dataset accounted for 80% of the views. This finding was echoed by Halvey and Keane (2007) who noted that 90% of videos in their data sample had few views. Silva and Dix (2007) also suggest user participation on YouTube follows the 1% rule: for every 100 people online just 1 will create content, 10 will interact with it, and 89 will only view it (Wikipedia, 2014). Halvey and Keane (2007) found that videos with higher views had more social interaction from users, evidenced by an increase in textual data and external links to the video. Users are more likely to comment on or share a video that is already popular. User information in their data sample suggested that users on average upload 2.5 videos and watch 456 videos. However, they noted considerable variance between users, with a few being more prolific than the majority. Marchionini *et al.* (2009) observed that most users only upload one video. Cha *et al.* (2007) found that rated videos accounted for only 0.22% of total views, and videos with comments only 0.16%. As well as commenting on videos, YouTube users interact by posting video responses to another user's video or their own previous videos. Rotman and Preece (2010) found evidence of small, special interest communities actively communicating through comments and video responses. Paolillo and Penumarthy (2007) also found evidence of social groups sharing niche content. They suggest that looking at social groups and the textual data

they produce could highlight less popular, more specific interest content, as the most popular videos on YouTube did not appear in their dataset. Tao *et al.* (2012) observed that in their small study of 440 videos, all relevant to a highly specific topic, the most relevant videos had received more user comments. Shamma *et al.* (2007) suggest there is no social network on YouTube but rather the videos are shared on other social networks for likeminded individuals to follow and watch. Broxton *et al.* (2011) measure the 'socialness' of a video by counting the number of external links. They found that 'socialness' does not correlate to long term popularity. However, Cha *et al.* (2007) found that the most popular videos in their dataset were linked; 47% of videos in their sample (252,255 videos from the Science and Technology YouTube category) had incoming links from external sites. Li *et al.* (2012) looked at how videos were shared over social networks and found that 40% of all YouTube views originate on Facebook. As of January 2011, 58.6 million videos were shared by Facebook users. The authors suggest that recommendation through sharing is a good way to find niche content over keyword search due to the lack of textual descriptions. Despite finding high numbers of external links to videos, Cha *et al.* (2007) observed that clicks from these links only accounted for 3% of the total views. Cheng *et al.* (2008) propose that users are more likely to browse recommended videos within YouTube than follow social links. Figueiredo *et al.* (2011) noted that users looking for specific interest videos are more likely to use YouTube's search facility, which as a text based search is inadequate to fully satisfy their search needs. Tao *et al.* (2012) explain that using text based search for a user generated video collection means that user queries need to match the user generated textual data (e.g., title, description, comments). Whilst this could be beneficial as both will use natural language, it creates a problem of ambiguity of terms (e.g., spelling mistakes, plurals, synonym use). There is also the problem that the text users enter may not adequately describe the video content.

Figueiredo *et al.* (2011) found the key source of video traffic on YouTube was from search and internal recommendation metrics.

2.2.2 Video Search

Two methods of video retrieval are defined by de Rooij *et al.* (2008): 'targeted' – search and 'exploratory' – browse. They advocate that a system which combines the two methods is optimal. Cunningham and Nichols (2008) found browsing to be more significant in finding popular videos than searching. 66% of YouTube sessions begin with a search, then users browse related videos to find subsequent videos rather than refining their search, suggesting that the majority of users have vague search needs. Tjondronegoro *et al.* (2009) discovered that users were more likely to click on a thumbnail of an image than a thumbnail still of a video, implying that a thumbnail still is insufficient for satisfying the search need. Kofler *et al.* (2012) indicate that query dissatisfaction happens because not enough information is presented to the user and not enough information is available to accurately index the videos. The authors noted low success rates in their experiments, with 36% of queries gaining no click-throughs. One still of a video may not show the section of video that is relevant; a user is unlikely to invest their time in a video if they are not certain that it is relevant. This causes a problem for designing an interface that allows the user to browse for videos and also for listing search results.

A plenitude of research exists that extends the success in improving image search and applies similar methods to try to improve video search. However, what has been successful for the still image does not transfer directly to the moving image. Goodrum (2003) discusses the problems facing video search. The multidimensional

nature of videos makes it difficult to fully describe the content. A video contains approximately 25 - 30 individual images or frames per second (each individual image is called a keyframe) and each keyframe contains both low-level features, including shapes, textures, colours, brightness, and high-level features, including people, places and things. Goodrum (2003) defines a 'shot' as: an unbroken sequential string of frames taken from a single camera. An extension of shot detection, 'scene' detection finds an object across the video but not necessarily in a continuous stream. The author points out that there is little research into understanding which features are most useful for search, ranking and categorisation. Three methods of video search are described by de Rooij *et al.* (2008) and Halvey and Jose (2012): Query by Text, Query by Example and Query by Concept. Query by Text, text-based search, relies on textual metadata and descriptions to index video. Using the same methods effectively applied to index web pages, text-based search retrieves videos based on text on the web page where the video is embedded. de Rooij *et al.* (2008) highlight the problem that video is not described well enough in text to be adequately indexed and retrieved in text-based search. Because video search is not precise enough, the retrieval process is slowed down by having to look through large quantities of results. Query by Example, content-based search, relies on the user uploading an existing video in order to find similar videos. Shot boundary detection methods are used to extract low-level visual features such as colour, shape and texture, and videos are grouped by visual similarity. As only low-level features are indexed, this search is very basic. McDonald and Tait (2003) found that users struggled to formulate search queries when using Query by Example search. Query by Concept, another content-based search method, groups videos based on semantic similarity. Semantic concepts are mapped to low-level visual features extracted from shots, scenes or keyframes. The majority of research focuses on applying image content detection algorithms to single shots of video (Morsillo *et al.*, 2010). Ulges *et al.* (2008a)

propose that training algorithms on multiple shots improves annotation performance. Query by Concept is not ideal until more high level semantic information is extracted from videos. However, only humans can extract high level features (Shih-Fu et al., 2007). Automatic concept detection methods need to be trained on existing textual data, but the textual data is insufficient. Until it is improved, automatic indexing will not be able to extract high-level semantic information.

Halvey and Jose (2012) argue that video search success is dependent on the precision of the query method and the ability of the user to specify their query. They compared the search methods of novice and expert users to establish whether having expert knowledge of video search improved users' satisfaction with search results. Their definition of novice user could be debated; 'Novice Users' were defined as computer science students and 'Expert Users' were defined as having knowledge of TRECVID. It is most likely that the novice users had some knowledge of social media and search methods, as opposed to a true novice internet user who perhaps needs help using Google. Regardless, they found no significant difference between the two user types. The authors suggest that video search is failing at the moment because it is not precise enough and users are not skilled enough at making successful queries. Video search at present cannot cope with specific searches and they propose that a combination of query methods is optimal for improving search.

2.2.3 Video Annotation – Bridging the Semantic Gap

Cha *et al.* (2007) discuss a bottleneck created by a lack of textual data that results in poor search and poor recommendation engines. This is defined by other authors as

the semantic gap. Enser (2008) describes the semantic gap as the distance between information that can be extracted automatically from a visual resource and how humans interpret the resource content. He argues that a description of the content of a visual resource lies in the semantic inferences represented in textual metadata rather than perceptual features. Perceptual features are indexed using content based search and textual data using text based search, neither method indexes the high level semantics required for image or video search. The information that can be retrieved from low level features cannot be transformed to high level features that represent objects in the image or video also, users formulate queries using high level semantics not low level features so what can be retrieved does not match what is being queried (Hare et al., 2006). Enser *et al.* (2007) found that the majority of terms people use to describe an image were not present in the image indicating a practice of subjective interpretation of the content. Tjondronegoro *et al.* (2009) describe bridging the gap between low-level visual features and high level semantic text. Bai *et al.* (2008) suggest mapping high level semantic concepts to low-level features, e.g., celebrity name to person. Enser (2008) states that at time of writing most attempts at bridging the semantic gap had not successfully addressed the problem of distance between object labelling and high level reasoning.

The most active research in this area is focused on trying to train algorithms to extract high-level features from online video in order to improve the precision and recall of video search. Morsillo et al. (2010) emphasise that current methods of automatic indexing are not transferrable to the web and large scale corpora of UGC such as YouTube. The majority of concept detection algorithms are trained on small, professionally annotated corpora, predominantly TRECVID, whereas YouTube is a large corpus and is user annotated. TRECVID (Smeaton et al., 2009) is a collection of

professionally annotated videos, mainly from the news genre, but increasingly since 2010 from other media outlets, namely the BBC, so the dataset more closely resembles web video (Over, 2014). An alternative dataset has been proposed which tries to more closely emulate video content that would be found on YouTube. Loui *et al.* (2007) created a benchmark dataset 'Kodak consumer video benchmark dataset' of annotated UGC videos. Videos are categorised by semantic concepts. There are two datasets: one containing videos uploaded by participants in a Kodak user study (1358 videos) and one of YouTube videos (4539 videos). Videos are annotated with predefined concepts rather than free natural language tags. The authors create an ontology consisting of seven categories, with 25 concepts for each category. However, little research has been published that uses the Kodak dataset. The TRECVID dataset annotates individual shots, whereas YouTube annotations tend to refer to the whole video.

Morsillo *et al.*'s (2010) experiments with YouTube, whilst offering some success, still only generated basic level vocabulary and at great computational cost, which is inappropriate for a home user. They acknowledge that video is more difficult to index by concept detection, as many single shots make up one video and content comes from audio as well as visuals. Jaimes *et al.* (2003) used speech from videos to create keywords to enhance the low-level visual features automatically extracted. Keywords are grouped into perceptual concepts based on the five senses. Min *et al.* (2003) also discuss a method of turning the audio commentary of a video into searchable keywords. Whilst this method is useful for extracting high-level semantic concepts, the problem lies in how reliable the transcribing software is. Ulges *et al.* (2008a) propose a system that learns from the low quality data available from YouTube. Although they improved annotations for a selection of videos, the

annotations were still at a basic descriptive level. Their approach is to use these methods to enhance textual data for existing text-based search rather than to categorise videos in semantic categories for content-based search. Despite research into content-based video search, the most popular method for users to find video online is using text-based search (Halvey and Jose, 2012), yet all research agrees that Query by Text is currently inadequate because of insufficient textual data and poor descriptions associated with online video. What is not agreed is which method should be used either to replace query by text or to improve it.

With users as content producers as well as content consumers, vast quantities of videos are being published with no editorial control. There is no control over metadata resulting in poor labelling and inadequate descriptions (Morsillo *et al.*, 2010). Bridging the semantic gap by creating improved annotations for videos is a lively research area, with a number of different approaches, both manual and automatic. Automatic methods concentrate on improving concept detection algorithms so they are able to extract high-level visual features and high-level semantic information. Manual methods look at employing or encouraging people to annotate videos.

Manual annotation is expensive and difficult to use for large scale repositories like YouTube. Shih-Fu *et al.* (2007), Tjondronegoro and Spink (2008), Ulges *et al.* (2008a) and Morsillo *et al.* (2010) argue that professionally annotated datasets like TRECVID (Smeaton *et al.*, 2009) are inadequate because the categories professionals use do not correspond to users' natural language used in search. The authors found that search terms used in YouTube did not correspond to the TRECVID semantic categories.

They argue that videos with bad metadata are invisible to users, which explains why the majority of videos on YouTube are difficult to find. Just because PGC is professional content does not mean it is professionally annotated. Although Halvey and Keane (2007) found that promoted videos have more descriptive information, the dominance of PGC in YouTube is a result of promotion rather than improved description or textual data. Higher quantity does not necessarily equal higher quality. There is to date no research that analyses the semantic vocabulary of this textual data to ascertain whether it is of an adequate quality. PGC might not fully meet the users search goal, yet videos that could satisfy their requirements are described poorly and therefore remain unfound.

2.2.4 How Users Search for Videos

There is a lack of understanding around how and why people search for videos on YouTube, the language they use and how successful search is for finding relevant videos. The knowledge that is available is based on web search for multimedia files. Vallet *et al.* (2008) specify an average of 2.21 search terms per query and Cunningham and Nichols (2008) found an average of 2.39 terms per query. Similarly, Tjondronegoro *et al.* (2009) discovered that the majority of video search queries are less than 4 words. The most popular searches were for people, celebrities, places or things with 64% of queries being for a specific person. Natural language searches received few click throughs, notably because of the lack of textual descriptions. Vallet *et al.* (2008) and Kofler *et al.* (2012) found that users use basic vocabulary for search terms and search for general subjects rather than specific interests, and that queries are unlikely to be repeated, reporting 57% and 36% unique queries respectively. In contrast, Conduit and Rafferty (2007) discovered that most people search for images using a mixture of both general and specific terms that describe people and things in

the image. Hollink *et al.* (2004) found differences in the way that users search for images compared to how they describe them. Ransom and Rafferty (2011) grouped individual index descriptions and search terms for images into general, specific and abstract facets. They found that users describe images with basic features and use more specific terms to search. Gligorov *et al.* (2013) analysed one month of query logs for an internal search engine on a Dutch TV company website, finding that approximately two thirds of the queries in their sample were unique. Kofler *et al.* (2012) infer that users simplify their queries to increase their chances of success as specific queries retrieve too few relevant results. They, Cunningham and Nichols (2008) and Tjondronegoro *et al.* (2009) found no evidence of users refining their queries. Marchionini *et al.* (2009) analysed their own queries used to harvest a collection of videos and retrieved fewer videos for highly specific queries. When analysing their queries with tags associated with the retrieved videos, they found little agreement on terms for specific interest videos and high agreement on terms for general topic videos. These results are indicative of the search problem for users. Rafferty and Hilderley (2005) discuss the difficulty of categorising and retrieving visual resources when there is no single multimedia indexing standard in common use. Without accurate textual descriptions for videos it is difficult for users to satisfy anything other than a very basic search need.

de Rooij *et al.* (2008) highlight the difficulty users have in simplifying their queries into terms that describe low-level features. It is difficult to reduce vocabulary to a basic level when users have a specific search goal in mind, for example searching for 'woman' when you want videos of 'Beyonce', or searching for 'dogs' when you want videos of 'poodles'. In this instance, users will resort to browsing and recommendation engines as they have reduced expectations of video search. The

other problem is agreement on terms. If the majority of search terms are unique and if the textual data available is only provided by the content owner, it is unlikely that many of the terms used will match the search terms of users. Users are familiar with text-based search, natural language and asking questions, however, there is not enough descriptive text available for videos. Rafferty and Hilderley (2005) refer to the difficulty of assigning textual data to visual resources when the meaning of the image or video must be interpreted, which is a subjective process. Language used in the interpretation needs to be shared by the indexer and the searcher. There is a need to understand the cultural context of the visual resource in order to successfully interpret its meaning. Rafferty (2011) claims that users will interpret non-textual information differently depending on cultural context and their own subject specific knowledge relevant to the time and place that they receive the information. As a result the words used to describe an image or video might change over time as cultural context changes. When content is interpreted rather than described accuracy is difficult to measure as many different meanings can be created. Successful interpretation will provide agreement on terms either between viewers or with the content creator as long as these interpretations are available as textual data. In terms of image or video retrieval, a unique interpretation would lead to an unsuccessful search.

Tao *et al.* (2012) measure the precision and recall of queries in YouTube using 17 terms containing plurals and synonyms to find videos about smokeless tobacco. The authors discuss the limitations of YouTube's text-based search facility and restricted filtering. Because YouTube search is text-based, user queries need to match user generated text data in titles, comments, descriptions, etc. They found a problem with the ambiguity of terms: plurals and spelling errors in the user generated data affected

the number of relevant videos retrieved. They suggest that multiple search terms are required to find relevant videos. If the average number of terms used in a video search query is less than 4 (Tjondronegoro et al., 2009), this indicates that users are unlikely to be satisfied by their YouTube searches. Similar to de Rooij *et al.* (2008), Tao *et al.* (2012) note the difficulty in formulating queries. They discovered that inconsistencies in their results, including plurals and synonyms could sometimes make little difference to precision and recall measurements, but that sometimes it could have an effect of up to 50%. Tagging, user generated metadata that is notoriously ripe with ambiguity, could increase the likelihood of matching search terms with textual data (Ransom and Rafferty, 2011) but, whilst it could provide the pool of text data required to increase precision, what cost would it have on performance (Gligorov et al., 2013)? de Rooij et al. (2008) emphasise the problem that video is not described well enough in text to be adequately indexed and retrieved in text-based search. Because video search is not precise enough, the retrieval process is slowed down by having to look through large quantities of results.

2.3 Games With A Purpose

Games With A Purpose (GWAP) were proposed with an initial focus on the ESP Game by Von Ahn and Dabbish (2004), which was then extended to further define the concept (Von Ahn, 2005; Von Ahn, 2006; Von Ahn et al., 2006; Von Ahn and Dabbish, 2008). The predominant theory of GWAP is to harness the man hours spent by people frivolously playing computer games and turn it into real work. In the US, 200 million hours are spent each day playing computer games (Von Ahn and Dabbish, 2008). The purpose of GWAP is to use humans to solve computational problems whilst entertaining them. The primary task that GWAP are used for is labelling web resources to improve indexing, search and accessibility. The majority

are developed from the design framework of the ESP Game. Von Ahn and Dabbish (2008) outline the key game elements of a GWAP: *Timed Response, Score Keeping, Player Skill Levels, High Score Lists, Randomness*. Two player games are recommended where points are awarded based on user agreement; in the ESP Game players enter tags to describe images, scoring points when either player enters a tag that matches with the other player. Taboo words are used in the ESP Game to provoke use of broader vocabulary. Once a tag has reached a certain threshold of agreement it is made visible to the players as a taboo word. The purpose of the ESP Game, to improve textual descriptions of images, was originally hidden from players. Users were asked to enter words to try to guess what their opponent is thinking.

GWAP have been developed further to encourage participants in Citizen Science projects (Cohn, 2008; Iacovides et al., 2013). GWAP use the concept of crowdsourcing, but rather than paying the crowd for their work, the crowd is rewarded through entertainment. Von Ahn and Dabbish (2008) describe three types of GWAP: output agreement, input agreement and inversion-problem games. Here, the output is the user input (i.e., tags) and the input is the web resource (i.e., image, video, audio or webpage). Inversion-problem games provide two player roles, describer and guesser and their aim is to elicit sentences or full strings that describe a resource rather than a section of tags. Input agreement games are seen by Law and Von Ahn (2009) as a way to avoid cheating and malicious data. The purpose of the ESP Game was to create labels that would improve image search and accessibility of the web for the visually impaired. However, Von Ahn and Dabbish (2004) describe limitations for improving accessibility because the game only generates lists of words rather than sentences that describe the content. Whilst useful, an improved method would be to encourage users to generate full sentences or query strings. Phetch is an

example of an inversion problem game and was created by Law and Von Ahn (2009). One player describes an image and the other player searches for the image via a custom search engine using the ESP Game dataset. Points are scored if the guesser finds the describer's image. A portal of GWAPs² was eventually created by a team of developers and researchers, with input from Von Ahn and Dabbish (2004), the ESP Game, Phetch, TagATune, Peekaboom, PopVideo and Verboosity (Law and Von Ahn, 2009; Von Ahn et al., 2006b). This portal has since been removed.

Goh *et al.* (2011) define two types of GWAP: Competitive and Collaborative. The ESP Game is an example of a competitive GWAP, where players work against each other to fulfil goal objectives; in a collaborative GWAP users work together. Goh *et al.* (2010) argue that collaborative GWAPs can incite malicious behaviour because the game is more open to useless data when competition to match with another player is removed. To avoid this problem, Ho *et al.* (2009) propose Kiss Kiss Ban, an image tagging game using the ESP Game datasets that incorporates both collaborative and competitive gameplay. Two players (guessers) collaborate to enter tags and try to reach agreement. They compete against one player (blocker) who has to prevent them reaching agreement by creating taboo words. The blocker and guessers compete against each other to score points for the round. Guessers do not see the taboo words, unlike in the ESP game. They propose that guessers will enter more specific terms while trying to avoid the blocker. The game was deployed on Mechanical Turk, so users were paid to play. It was played 537 times generating 5,321 labels. The ESP game data was used as ground truth data and 78.84% of labels

² <http://www.gwap.com>

generated matched with the ESP game labels. As users were paid to participate, it is impossible to accurately assess how users were motivated by the change in gameplay. Users are motivated differently by financial incentives than by an incentive to be entertained (Melenhorst and van Velsen, 2010).

The majority of GWAP research concentrates on implementation and little work has focused on assessing tag quality or user perception of GWAPs. Because most methods are based on replicating Von Ahn's methods (Von Ahn and Dabbish, 2004; Von Ahn, 2005; Von Ahn, 2006; Von Ahn and Dabbish, 2008) the general consensus is that tag quality is measured by user agreement; very little research has been conducted to evaluate the quality of tags, with exception Goh *et al.* (2011) and Gligorov (2012). The ESP Game provides a benchmark standard from which the majority of GWAPs are built. However Ma *et al.* (2009) warn that not all tasks can be made fun by simply applying the ESP Game model. Robertson *et al.* (2009) question how useful tagging GWAPs are when following the ESP Game template, if they only generate low-level basic tags (when automatic methods are capable of this) and how effective these tags will be for improving search if only the quantity of general descriptions is increased. Von Ahn and Dabbish (2004) and Gligorov *et al.* (2013) evaluated the usefulness of the tags generated by performing search tasks using the tags as keywords in custom search engines. Both authors found that the tags achieved high precision results. Sheng *et al.* (2008) suggest that many labels, even if noisy, are better for data mining than few labels or none.

Goh *et al.* (2010) compare the effectiveness of collaborative and competitive GWAPs using two small-scale image tagging games, with a non-game system being created as a control. Players were recruited rather than being freely attracted online. The authors created ground truth data consisting of 20 tags per image that could be used

to measure the usefulness of user tags. They proposed two measures of usefulness, *accuracy* and *diversity*. Accuracy is the mean number of tags that agree with the ground truth data, while diversity is the mean number of unique tags. Users expressed a preference for the competitive game over the collaborative game although this was not significant. User preference for game over non-game system was significant. The competitive game produced more unique tags, but also more matching tags. The lowest number of tags was generated in the non-game system. Tags were assigned to three levels: Level 1 General, Level 2 Specific and Level 3 Highly Specific as proposed by Golder and Huberman (2005). Goh *et al.* (2011) found more Level 1 tags in the non-game and the competitive GWAP and more Level 2 specific tags in the collaborative GWAP. They found very few Level 3 tags in either context. User studies highlighted that players preferred a game-based tagging system to a non-game tagging system. Goh *et al.* (2011) note that game-based tagging is the most suitable approach for generating user data, but that the challenge is in creating engaging competitive games.

To create a sense of community to encourage participation, Rafelsberger and Scharl (2009) deployed Sentiment Quiz on Facebook, a method also applied by Barrington *et al.* (2009) with Herd It, and Poesio *et al.* (2013) with Phrase Detectives. Rafelsberger and Scharl (2009) hypothesise that if GWAPs target a specific community, users will have a high level of intrinsic motivation to play to help build a shared knowledge repository. During a 3 month period, more than 1000 Facebook users played. In contrast Poesio *et al.* (2013) found that more users came from direct links than from Facebook integration. Lin *et al.* (2008) claim that the number of players willing to invest time and effort in a GWAP is limited. In relation to this, Poesio *et al.* (2013) claim that a good advertising strategy was imperative to the success of Phrase

Detectives. The GWAP ran for 3 years, during which 8000 players were attracted. They performed 5000 hours of work equal to 2.5 person years and entered on average 450 annotations per hour. Average time playing the game was 35 minutes and 5 seconds. 46% of their traffic came through direct hits, 29% through web links, 13% through Facebook advertising and 12% through search. The bounce rate (the percentage of users who leave the website without clicking any links) was high 33% for direct hits, 29% for web, 44% for search and a massive 90% for Facebook, proving that Facebook advertising did not work.

Along with the ESP Game another seminal piece of research that inspired the area of GWAP is the Steve.Museum project (Trant and Wyman, 2006), a tagging game that was created to encourage the general public to tag pieces of art. The purpose was to bridge the gap between how experts and the general public describe art, to improve access to collections. 1,313 works of art were tagged with 7,339 tags between October 2005 and September 2008. The majority of tags, 90.2%, provided new, additional metadata that could be used to support professional descriptions.

Von Ahn and Dabbish (2008) recommend a set of metrics that can be used to measure the efficiency of a GWAP:

- *Throughput* – the amount of resources tagged in a human hour. A human hour is calculated as the average of all game sessions over a set time period for all players.
- *Average Lifetime play (ALP)* – the amount of time that a game is played by each player averaged across all people who played it.

- *Expected contribution* – throughput multiplied by ALP.

Of the researchers that have employed this method, the average number of tags per minute assigned to a GWAP is between 3 and 4 (Barrington et al., 2009; Von Ahn and Dabbish, 2004; Ho et al., 2009).

A further method to measure the success of a GWAP is proposed by Law and Von Ahn (2009): the *Player Retention Curve* is a measure of play frequency. The number of games each player played is ranked by frequency and then plotted on a log-log scale. The angle of the curve reveals player patterns. A steep curve indicates many people played a few times before abandoning the game never to play again, and a flat curve indicates that many people played a large number of games. When usage of GWAP.com was measured, they found that few players played many games, and that most players played few games. There was little distinction between the curves for all games. Goh *et al.* (2011) discuss barriers to user enjoyment and output quality:

- *Gameplay* – competition, scoring only upon agreement with another player, encourages basic level tags as there is more chance of agreement if using general, obvious descriptive terms.
- *Time Limit* – placing a time limit on players increases competition but also feelings of frustration. Typing speed can be a barrier to the game. It is easier to think of a basic level tag under a time limit. Users spend less time thinking about tags in a game environment.
- *Scoring* – pressure to score in a fixed amount of time induces frustration that may result in poorer quality tags and reduce user enjoyment.

Goh *et al.* (2011) propose seven characteristics for enjoyment of a GWAP that can be measured in a playability questionnaire: Appeal, Challenge, Usefulness, Absorption, Control, Learnability and Social Interaction. To optimise user performance of a GWAP to categorise pieces of music, Barrington *et al.* (2009) suggest a user centred design approach, creating prototypes and observing players as well as conducting usability questionnaires. They propose four methods for engaging players in a GWAP:

1. Create visually appealing and intuitive interfaces;
2. Create genre specific games;
3. Create competitive scoring and real-time feedback;
4. Create community by integrating with Facebook.

Despite these design decisions, they observed that users were more interested in the musical content of a game than in the gameplay. Von Ahn and Dabbish (2008) suggest that being told to label an image would have deterred players. "People will play such games to be entertained, not to solve a problem—no matter how laudable the objective." Von Ahn (2006). The ESP Game's success lies in the fact that users were asked to guess what the other player was thinking. This became the goal and players were unaware of the purpose behind their actions. Goh *et al.* (2011) found users more likely to play a game if they can see why it is useful. Iacovides *et al.* (2013) discovered that the majority of users played because they wanted to contribute to the project. They suggest providing regular progress reports on how users' contributions are helping or letting users see the benefits of their contribution (van Velsen and Melenhorst, 2009) to help prolong use.

The cold start problem is a difficulty for GWAPs because the lack of existing data means there is nothing to compare the generated tags to. Gameplay is also affected as scoring typically relies on user agreement; the two player matching format of the ESP Game avoids the problem only if two players play in real time. Ground truth or gold standard data is used in many GWAP projects to counteract the cold start problem and as a benchmark to evaluate tag relevance and quality assessments. This data is produced by either the research team, professional annotators, dictionary integration, automatic methods or existing benchmark data such as the ESP Game or TRECVID. The latter methods restrict content in the GWAP (Law and Von Ahn, 2009; Poesio et al., 2013). Chamberlain *et al.* (2009) suggest four things that can affect tag quality: misunderstanding the task, attention slips, malicious behaviour and genuine ambiguity of data. They found high agreement between expert gold standard data and user data, a finding echoed by Gligorov *et al.* (2010) and Rafelsberger and Scharl (2009). This is in contrast to Von Ahn and Dabbish (2008) who claim that professional annotations are of higher quality. The main challenge for GWAP designers is that quality of output is as important as user enjoyment. If the GWAP is not playable, users will not enter data but equally, the data generated needs to be of high quality and useful for the purpose (Rafelsberger and Scharl, 2009).

For video tagging games, the data has to be relevant to the video and describe the content. A weakness in the proposals by Von Ahn (Von Ahn and Dabbish, 2004; Von Ahn, 2005; Von Ahn, 2006; Von Ahn et al., 2006a; Von Ahn and Dabbish, 2008) is the lack of attention given to assessing the quality of the user inputted data. They define tag quality as whether there is user agreement above a threshold of 2. This dismisses a large proportion of unique tags that could accurately describe content. Law and Von Ahn (2009) define verified tags as a tag that has agreement, a measure applied

by Gligorov *et al.* (2010). They claim that agreement is not the best measure for relevance as non-verified tags are also relevant for search, in particular when describing audio content as this provokes more descriptive and subjective language. However, they provide no alternative method to assess tag quality. Game elements in the ESP Game encourage users to enter basic level tags, describing the most obvious aspects of the image. Goh *et al.* (2010) asserts that the output agreement method encourages users to enter more generic descriptions. Robertson *et al.* (2009) recognise the requirement for GWAPs to encourage more specific descriptions and suggest removing the emphasis on user agreement in order to achieve this. To test their theories, they created a robot to play the ESP Game. The robot managed to score points and reach agreement quickly by entering synonyms of the taboo words. Ho *et al.* (2009) and Robertson *et al.* (2009) highlight the fact that visible taboo words encourage users to enter synonyms of the taboo words. Ho *et al.* (2009) propose not presenting taboo words to users and enforcing penalties if they are guessed, encouraging users to enter more specific terms while evading the restriction. Jain and Parkes (2009) also emphasise the requirement to concentrate gameplay on rare words first and less on early agreement. Robertson *et al.* (2009) note that the longer a user tags, the less obvious and more specific their tags become. The difficulty lies in determining how long a user can tag for before the task has a negative impact on enjoyment. Thaler *et al.* (2011) provide a useful survey of GWAPs, there are too many individual projects to discuss individually here. Video tagging games will be discussed in the next section.

2.3.1 Video Tagging Games

Tagging a video takes high mental focus, (Wang *et al.*, 2012). Users could be frustrated by the intensive labour cost of manual tagging, which could limit both the

quantity and quality of tags. One proposed method of improving user motivation and encouraging users to tag videos is to turn the tagging system into a game.

Whilst research into GWAP, and in particular image tagging games, is fairly extensive, there is very little research into video tagging games. The original version of VideoTag (Greenaway, 2007) was the first video tagging game. The VideoTag tagging experiment consisted of a one player game where users were encouraged to tag a selection of sixty carefully chosen, funny YouTube Videos. The small scale, proof of concept experiment was uncontrolled with random users being attracted to the game through promotion on various Web 2.0 sites. The makers of the ESP Game created the video tagging game Pop Video, but there has been no published research on their findings. Yahoo! made a brief foray into the video tagging area with Video Tag Game (van Zwol et al., 2008). Their small user study found a high level of agreement on tags and that users reached agreement quickly (within a few seconds). The research focussed on using user tags to retrieve fragments of video that are relevant to a search rather than the whole video. No further developments have been published for this project. Siorpaes and Hepp (2008) created OntoGame, a GWAP for crowdsourcing the creation of ontologies for classifying online video by semantic relevance as opposed to a video tagging game. Users were asked questions about the video and videos were classified into ontologies based on user agreement. Gomes *et al.* (2013a) created a GWAP for annotating the audio of movie clips to improve audio description.

One notable video tagging project has been Waisda?. Initiated by the Netherlands Institute for Sound and Vision it was created not to label online video but to improve

access to an existing archive. Gameplay, development and implementation are discussed by Hildebrand *et al.* (2013). Waisda? was developed over two pilot studies in 2009 and 2011, in collaboration with Dutch TV (Hildebrand *et al.*, 2013). Hildebrand *et al.* (2013) and Oomen *et al.* (2010) attribute the project's success to promotion by the Dutch TV companies. Differing usage statistics were reported for both studies by different authors, perhaps because data samples were taken at different stages in the project. In pilot one, which ran for 8 months 2,400 unique players tagged 650 videos (Lin and Aroyo, 2012) generating over 420,000 tags, an average of approximately 700 tags per video (Hildebrand *et al.*, 2013). In pilot two, 436,456 tags (Gligorov *et al.*, 2013) were assigned to 2500 videos by approximately 530 players over 3.5 months (Lin and Aroyo, 2012). Gligorov *et al.* (2013) report an average number of tags per video of 199; 44% of these tags were unique. Lin and Aroyo (2012) claim an improvement in efficiency of the tagging game in pilot two measured by a decrease in the number of users and a reduction in time taken to generate a similar quantity of tags.

Whilst the game followed the same format as the ESP Game and therefore the game itself was not new, this is the first tagging game research that analysed the quality and usefulness of the output data of the game. The tag analysis is discussed in Section 2.6.3. Gligorov *et al.* (2013) produced the first study that evaluated the search performance of tags generated through a GWAP. The search performance of 12 search engines that index different types of metadata associated with videos was evaluated. Combinations of either all user tags, unique tags excluding (verified tags), professional annotations and closed captions were tested. The authors measured search performance using the MAP (Mean Average Precision) measure. Search engines that index user tags outperformed other engines by 33% and search based on

user tags alone outperformed them by 5%, continuing to improve over time. Search engines which index verified tags only gave poorer performance, suggesting that a range of tags, not just tags on which people agree is best. It should not be assumed that only tags on which users agree are relevant to a video. Verified tags gave higher precision but had lower recall. Search using all user tags provided relevant results that were not found by the verified tags search. These findings are interesting as they go some way to proving that user tagging is a useful tool for improving text-based video search. Indexing user generated tags could yield better results than more expensive metadata such as professional annotation or automatic generated annotations. The authors note that filtering tags by user agreement eliminates many useful tags. Therefore finding a method to filter tags by relevance is required to improve precision without affecting recall.

Pop Video, Video Tagging Game and Waisda? all follow the same game format as the ESP Game: a two player game where points are scored when the tag you enter matches the tag of the other player. Points are added to a leaderboard, which is displayed at log in. These three games are collaborative GWAPs with competition only coming from the leaderboard. Pinto and Viana (2013) describe a one player video tagging game (TAG4VD) based on VideoTag version 1 (Greenaway, 2007). Like Waisda? and Video Tag Game, their research focuses on tagging fragments of video rather than the entire video like in VideoTag, similar to the benchmark dataset TRECVID (Smeaton et al., 2009). In Waisda? and TAG4VD tags are assigned to a timestamp in the video referred to as time tags or deep tags, rather than to the whole video (global tags). Lin and Aroyo (2012) claim that user inaccuracy could reduce tag relevance when assigning tags to time codes. Rather than scoring when a tag matches that of the opposing player, in TAG4VD same word tags are clustered based on

distance from the timestamp where the word was first entered, a technique introduced by Begelman et al. (2006). Points are scored based on how close in time the tag is entered to the timestamp of the cluster. In the absence of automatic methods to check if a word is relevant to the video (if this existed there would be no problem for video search) video tagging games use co-occurrence as a measurement for relevance. The problem with this method is that it can exclude valid tags of high specificity as user agreement tends to be better on general tags (Halpin et al., 2007; Golder and Huberman, 2006). In a separate interface and away from the gameplay, Pinto and Viana (2013) offer users the opportunity to assess the relevance of tags entered in the game by selecting whether a tag is good or bad, i.e., relevant or not relevant. The authors intend to discover if tagging can be used to highlight the most interesting fragments of a video. There are no published findings on TAG4VD.

Problems for video tagging games are primarily centred on motivation to play and quality of output. Unfortunately, little has been published to date that fully investigates either area. Wang *et al.* (2012) and Goh *et al.* (2011) claim the high cognitive cost to the user of thinking of terms that describe content in a limited period of time. This cost increases the more specific or abstract the terms are required to be. Lin and Aroyo (2012) consider the challenge involved in developing an interface that encourages high quality tagging when the video has to be the overriding feature.

2.3.2 Crowdsourcing and Motivation to Participate

Motivation to engage in a GWAP is multifaceted and dependent on the purpose of the game. One facet common to all is crowdsourcing. Dunn and Hedges (2012)

conducted a survey of participants actively involved in crowdsourcing projects in an endeavour to discover the reasons why they participate and the benefits of crowdsourcing projects over traditional methods of data gathering in the humanities. Crowdsourcing projects should allow a large number of users to be involved, a critical mass, even though only a few will actively engage. Attracting the 'critical mass' of users means to elicit enough participants that the system becomes self-sustaining, with its perceived popularity creating further growth. Hsu and Lu (2004) explain that a user's perception of critical mass is a motivation for participation, in other words, if a user perceives the game to be popular and thinks a lot of their peers are using it, they will also participate. Sweetser and Wyeth (2005) argue that social motivation can make people play a game they do not like or play a game when they do not like games at all. Dunn and Hedges (2012) suggest that engaging in a crowdsourcing project involves more effort than engaging with a social network, forum or playing a game. Users need a level of commitment to the project, which is difficult to incentivise. The authors record a finding that subject interest was the key reason for participation; interest in the subject gives users the desire to contribute. Oomen *et al.* (2010) noted that 70% of visitors were from external websites, enforcing the notion that extensive promotion with specific target groups is required to attract players. This questions how motivated the general public are to play video tagging games. Is their incentive more to watch videos they want to watch rather than the tagging activity or engaging in the game elements?

Oomen *et al.* (2010) claim that video content is a key motivational factor and acknowledge a need to keep content fresh; Von Ahn and Dabbish (2004) also emphasise that the content being tagged affects enjoyment. The most tagged videos in their experiment were fragments of a popular Dutch reality TV show watched by

millions. The authors found that users invited by a trusted organisation (in this case a Dutch TV channel) were more likely to play. This finding is echoed by Dunn and Hedges (2012) and by the Your Paintings project³ initiated by the Public Catalogue Foundation in partnership with the BBC and funded by the Arts council of England. At the time of writing, 10,184 taggers have entered 3,404,673 tags for 188,644 paintings. The ESP Game was constantly advertised on TV and in the press (Poesio *et al.* 2013). Google invested in the ESP Game and commissioned it for their image labeler. This project had phenomenal success and influenced future research, including this thesis. Von Ahn and Dabbish (2004) reported that over a four-month period from August-December 2003, 13,630 people generated 1,271,451 tags with 33 people playing more than 1000 games. The success of the ESP Game has never been replicated, suggesting that a large incentive to participate was in the novelty of the idea. It also indicates that involvement and promotion by trusted organizations motivates users to play the games more than the activity itself. Law and Von Ahn (2009) compared user activity in music tagging GWAPs (Turnbull *et al.*, 2007; Mandel and Ellis, 2008; Kim *et al.*, 2008). They found that their TagATune GWAP considerably outperformed the others by almost 14,000 users. Whilst the authors credit their game methods for this, it is pertinent to speculate whether TagATune's deployment on the GWAP.com website had more impact on usage than game quality.

Shin and Shin (2011) suggest that the greater the trust a user has in a game the greater their intention to play. Salen and Zimmerman (2004) affirm that users must

³ <http://tagger.thepcf.org.uk/>

feel a sense of safety and trust in order to play. This is external to the game and comes from recommendation, either social or via organisations. Hsu and Lu (2004) argue that participation is not driven by a user's perception of the project's purpose but simply by their perception of how popular it is. Poesio *et al.* (2013) argue that building attractive games is not enough to attract players, it is also necessary to develop an effective advertising strategy. In a saturated casual games market, competing for attention by constant promotion is essential. Achieving visibility and maintaining it requires constant effort. Their Phrase Detectives game was featured in numerous blogs, The Times, BBC, gaming forums and a pay per click advertising campaign was initiated on Facebook. Dunn and Hedges (2012) argue that the most successful projects elicit participation from interested and engaged members of the public, noting that mass media exposure to projects leads to a spike in user activity. The authors found no single motivation to participate; both personal and social motivations were present, although most users were motivated by a benefit to help others (altruism) rather than by personal gain. Oomen *et al.* (2010) also identified altruism (the selfless act of investing time in a project without reward) as a key motivator for players of Waisda?. Will players only be motivated by altruism if they are directed to the game by an organization they trust and have an interest in? Mekler *et al.* (2013) created a simple image tagging GWAP to test whether users were motivated to participate more by points or by purpose – the purpose being to benefit computer understanding of images. The authors found that motivation to participate was improved if either or both conditions were present. Whilst users in the points condition generated more tags, users motivated by purpose produced better tags. Assigning meaning to the task encouraged users to spend more time on the task. Goh *et al.* (2011) also highlight altruism as a key incentive that encourages tags of high quality.

Whilst Oomen *et al.* (2010), Dunn and Hedges (2012) and Hsu and Lu (2004) identify the need to reach out to a 'critical mass' of users, the aim should not be to reach the widest audience but to attract 'super taggers'. Super taggers defined by Trant (2009) are the few users who provide most of the tags. Taggers follow the fabled 80/20 rule: whereas the majority of users will enter one tag, a few super taggers will enter thousands. Kuittinen *et al.* (2007) found that 98% of users will play out of a motivation to 'just try it out', also reported by Trant (2006). Hildebrand *et al.* (2013) indicate a requirement to reward super taggers, but what motivates a super tagger? Super taggers could be motivated by a keen interest in the process of tagging, interest in the subject matter they are being asked to tag, an affiliation with the organisation/creator of the system or motivation to help the project (altruism). It could be assumed that in earlier iterations of tagging systems during the Web 2.0 boom of the mid 2000s there was more likelihood of finding super taggers primarily motivated by the tagging activity. Individual projects acknowledge super taggers or power users, Dunn and Hedges (2012) conducted a survey with recognised super contributors from a selection of Citizen Science projects. They observed a strong sense of competition between the users that was missing in a survey of regular contributors. The authors note that the sense of competition was not necessarily derived from competing against other players, but in feeling they had performed well, the desire to create the most or the best quality data and a feeling of success. Arends *et al.* (2012) found 8 super taggers out of 182 users, with the super taggers entering 91.1% of the tags. Chamberlain *et al.* (2012) found that the 10 highest scoring players (1.3% of all players) made 73% of the annotations. In a Facebook version of their GWAP the 10 highest scoring players (1.6% of all players) made 89% of the annotations. When comparing male and female power users, they discovered that females made more annotations than males, 48,359 versus 4,817 respectively.

Understanding the potential audience for the GWAP and promoting it accordingly could increase the likelihood of attracting power users.

Methods to improve user motivation to sustain play are proposed by Oomen *et al.* (2010) but remain untested. They suggest letting users browse the tags that have been entered to find other videos to watch. Letting users see how the tags can be used and see that they are useful creates a new motivation, content discovery. This facility is provided by Pinto and Viana (2013), but there are no published findings as to whether this has actively motivated users. Another suggestion is to allow users to unlock video content the more they play, giving the incentive to play for longer and return to the game, this also works to keep content fresh. Poesio *et al.* (2013) suggest allowing users to give feedback on the game and offer suggestions. Ryan and Deci (2000) emphasize that autonomy, the ability to make choices, is synonymous with greater engagement and better performance. Offering the players the opportunity to affect the game through feedback could increase motivation by improving their sense of affiliation with the game.

Chamberlain *et al.* (2012) describe three incentive structures for participation in a GWAP: *personal*, *social* and *financial*. Personal incentives are contribution to the project, interest in the subject matter or desire to improve at the task. Social incentives are to compete against peers, by scoring more points or reaching higher levels. Financial incentives, the reward for effort with money, are an extrinsic motivation; financial incentives have a negative impact on intrinsic motivation and can deter a user from participating (Melenhorst and van Velsen, 2010). Chamberlain *et al.* (2012) hypothesise that a combination of the three types is essential for

sustained play, but the majority of players never played past level two. Most users tried the game then never returned, a motivation defined by Trant and Wyman (2006) as *Just Try It Out* and by Ryan and Deci (2000) as the intrinsic motivation of curiosity, to take interest in novelty. Furthermore, social posts from within the GWAP were related to content rather than player activity in the game. This would imply that users were not motivated by level progression or point scoring but by the subject matter (also shown by Stuart (2012), Arends *et al.* (2012) and van Velsen and Melenhorst (2009)) and personal incentives to complete the task.

2.3.3 Theory of Play - Motivation, Engagement and Flow

Ryan and Deci (2000) claim that to be motivated means to be moved to do something. The level of motivation a user has to complete a task is dependent on the type of motivation. There are two types of motivation: *intrinsic (IM)*, doing something because it is interesting and enjoyable and *extrinsic (EM)*, doing something because it leads to an outcome. Ryan and Deci (2000) define a third state of motivation as *Amotivation (AM)*, which is not being compelled to act. Vallerand (1997) suggests three cognitive levels of motivation: *global*, *contextual* and *situational*. Situational motivation refers to the activity or task, therefore all further discussion of motivation will refer to motivation at the situational level. Intrinsic motivation is predominately personal; a user will engage in a GWAP out of a desire to partake in a fun or challenging activity. They will be interested in the subject matter, be curious and interested in the novelty. Voiskounsky *et al.* (2004) recorded curiosity as the main motivation to play a game. Users will be motivated to complete the tasks for the personal achievement of success irrespective of any reward. IM exists only in the relationship between user and the system and it will differ from person to person. IM is catalysed rather than caused. Tasks within the GWAP can elicit, sustain or enhance

IM to facilitate it, or subdue or diminish IM to undermine it. A user must feel that their actions are at their own volition and not controlled in any way by the GWAP; any rewards, social interaction or feedback must be conducive to feelings of competence in order to maintain high levels of IM. Extrinsic rewards can undermine IM, whereas competition pressure as long as the user feels they have choice and self-direction can enhance it (Ryan and Deci, 2000). Extrinsic motivation is predominately social, consisting of external influences that make a user complete a task. EM is important if a task is dull or arduous, to encourage users to do it. Two types of EM can be defined when Self Determination Theory is applied (Ryan and Deci, 2000), *self-determined extrinsic* and *non-self-determined extrinsic*. They differ by autonomy; with self-determined extrinsic the user freely chooses to engage, with non-self-determined extrinsic the user feels controlled by an external force to undertake a task.

Ryan and Deci (2000) explain that motivation can be ordered by the extent to which the behaviour emanates from within, where *amotivation* is the most external and *intrinsic* the most internal. They further divide EM into four categories, *external regulation*, *introjection*, *identification* and *integration*, external regulation being a step up from AM and integration a step down from IM. External regulation and introjection are classed as non-self-determined extrinsic, identification and integration are classed as self-determined extrinsic and intrinsic is also self-determined. Both *external regulation* and *introjection* are externally motivated states. *External regulation* describes participation in an activity purely for personal reward or to avoid punishment i.e. a child carrying out household chores to earn extra pocket money or doing the chores to avoid getting in trouble with their parents. The activity is often conducted begrudgingly and with little enjoyment or engagement. *Introjection* describes

participation in an activity for approval from the self or others to satisfy the ego i.e. doing extra work to impress a boss or working out relentlessly in the gym. In *identification* state, motivation starts to become internalised. People will participate if they value the activity or the activity helps another person they value. Goals are self-endorsed rather than receiving external reward i.e. shaving your head for charity (when praise or ego boost is not the goal) or completing a survey with no potential reward at the end. *Integration* describes a state where although the person was initially extrinsically motivated to participate, through enjoyment of the task motivation has been internalised so that the person is now more intrinsically motivated to participate. i.e. a person starting a job for the financial reward but putting in more effort and taking more pride in their work without expecting approval (*introjection*) or more pay (*external regulation*).

This scale can be used to measure the likelihood of a user engaging in a GWAP. In terms of participation in a GWAP, AM is having no interest at all. External regulation would be participation as part of a controlled experiment, for extra module credit, through Amazon's Mechanical Turk or for other financial reward. Introjection would be to 'try it out' out of a vague link to the creator or social pressure to be seen to have used it, and the user would feel some element of external control over their actions. With identification, a user would value the GWAP in some way, be that the concept, the purpose or the organisation and they are more likely to value the extrinsic rewards for contribution than gain personal satisfaction from contributing. Finally with integration, the users may have been attracted to the GWAP by extrinsic rewards but find a great affinity either with the purpose or the gameplay and the reward they feel for participating becomes more akin to that of an intrinsically motivated player. Quality of experience for the user and also quality of output is

improved if a user participates freely with an internal motivation. Intrinsically motivated users will produce higher quality output (Vallerand, 1997). Users attracted to a GWAP through IM will most likely become super contributors. The same is true for Integration EM.

If a user is attracted to the game through the intrinsic motivation of curiosity, there is no guarantee they will become a super contributor. Other factors must be present, namely flow. Flow as a state of immersion in an activity in everyday life was defined by Csikszentmihalyi (1975). Flow is defined as “the holistic sensation people feel when they act with total involvement” (Csikszentmihalyi, 1975, p.36). People become completely immersed in an activity to the point that they lose awareness of time and surroundings, all things but the activity itself. Flow cannot be achieved unless a person’s perceived skills match the perceived challenge. Sweetser and Wyeth (2005) claim that flow can only begin if a person feels their skills match the challenge set. The flow state emerges between the point of anxiety and the point of boredom. If a task is too difficult a user will feel anxious, if it is too easy a user will feel bored. There are eight elements to flow:

1. Balance between perceived skill and perceived challenge;
2. Clear goals;
3. Immediate feedback on actions;
4. Action and awareness are merged;
5. Able to fully concentrate on a task;
6. Sense of control over actions;
7. Loss of self-consciousness;
8. Sense of time distorted.

The first five are prerequisites, the last three describe the state of flow, and all eight need to occur for flow to exist. Ryan and Deci (2000) argue that motivation is determined by three elements: the user's perceptions of autonomy, competence and relatedness. An intrinsically motivated user will feel an association with the task (relatedness), feel able to complete the task (competence) and feel they have control over their actions (autonomy). Providing that the interface and gameplay support the user and facilitate their motivation, there is a high probability that an intrinsically motivated user will experience flow as the five pre-requisites are being met. Csikszentmihalyi (1975) refers to intrinsically motivated people as *autotelic*; they gain reward from participation in the activity itself rather than having any expectation of future benefit. He found that high intrinsic motivation is positively linked to high instances of flow. Mannell *et al.* (1988) contradict this finding, suggesting that extrinsically motivated users who freely choose to engage in an activity report the highest instances of flow. However, Kowal and Fortier (1999) argue that Mannell *et al.* (1988) did not take into account Self Determination Theory and the two types of extrinsic motivation. They found that SD extrinsic motivation and IM gave higher instances of flow than Non SD extrinsic motivation. Hsu *et al.* (2007) propose that game designers should emphasise aspects of the game that will attract intrinsically motivated users as opposed to extrinsically motivated ones to improve chances of creating a critical mass of players and flow experience.

Echoes of Csikszentmihalyi's Theory of Flow can be identified in theories of play. Huizinga (1949) wrote the seminal Theory of Play establishing seven characteristics of play:

1. Play is voluntary;
2. Play is outside ordinary life;

3. Play is not serious;
4. Play is utterly absorbing;
5. Play has no material interest or profit;
6. Play proceeds according to rules;
7. Play creates social groups separate from the outside world.

A person must have choice and clear goals, most perceive they can accomplish the activity and must feel control over their actions to execute a state of play or flow. Both play and flow are intrinsically motivated activities. A person freely engages in the activity because of the enjoyment it brings with no thought for any rewards the activity may bring. If extrinsically motivated, play stops being voluntary, it is controlled and is no longer play (Huizinga, 1949). As previously discussed, self-determined extrinsic motivation can become internalised. In this instance play and flow can still be present despite initial extrinsic motivation to participate in an activity. Play can also be taken too seriously so that play becomes work, (Huizinga, 1949) (e.g., professional sports people or musicians). Equally, serious activities can be approached in a playful manner (Huizinga, 1949); work can become play, making a task more enjoyable e.g., GWAPs (Fine, 1987). Huizinga (1949) asserts that “all play is meaningful”, as opposed to the antiquated viewpoint of some (put forth by Sutton-Smith (1997) as one of seven rhetoric’s of play) that play is a frivolous activity with little purpose. Salen and Zimmerman (2004) suggest that their theory of meaningful play is a prerequisite of flow. They offer two methods of describing meaningful play: *descriptive* which is the relationship between player action and player outcome, and *evaluative*, which is the emotional experience created by those actions. Evaluative play is further sectioned into two types, *discernible* and *integrated*. Discernible is the short term effects of actions and integrated is the long term effects. Discernible relates to the flow prerequisite of clear feedback and integrated describes how choices are integrated through goals, challenges and uncertainty. An analogy is presented by

(Sutton-Smith, 1986, p.185) warning of relying on flow to describe play “Flow is to play what orgasm is to sex.” Play can be meaningful without flow, the same as sex can be meaningful without orgasm. Cowley *et al.* (2008) surmise that flow and play are not mutually exclusive; play does not need to have flow to exist and flow does not always come from play, however, play will be sustained if the person enters flow state. The English language offers two separate words with two different meanings, *Play* and *Game*. This is not the case in other languages where *play* is the verb and *game* is the noun, for example in German: man spielt ein spiel - I played a game (Salen and Zimmerman, 2004). Whilst games are played, not all play activity is conducted during a game. A game exists without play. Play is a human condition, as is flow.

Huizinga (1949) acted as a catalyst for other researchers to develop theories of play focussing primarily on ludic behaviour. More recently research has begun to assert them toward the playing of games rather than play as ludic behaviour. This has not resulted in a single agreed upon definition of play, but in a number of overlapping characteristics of what play is. Caillois (1961) offers a definition of play in notable research that critically develops the theories of Huizinga (1949). Play is an activity that is free, separate, uncertain, unproductive, governed by rules and make-believe. A general definition of play: “Play is free movement within a more rigid structure” is proposed by (Salen and Zimmerman, 2004, p.311). They warn, however, that play and games cannot be fully understood by one single definition as they consist of many forms and concepts. Sutton-Smith (1997) argues that play cannot be summarised in a single definition. He suggests that ludic activities can be categorised as nine play forms, ordered from most personal to most public:

1. Mind or Subjective – fantasy, role-play;
2. Solitary – hobbies, exercise, collecting;

3. Playful Behaviour – tricks, pranks;
4. Informal Social Play – leisure activities e.g. parties, pubs, clubbing, theme parks;
5. Vicarious Audience Play – TV, Film;
6. Performance Play – instruments, acting;
7. Celebration/Festival/Ritual – Christmas, Halloween;
8. Contests – sports, competition;
9. Risky or Deep Play – extreme sports.

Sutton-Smith (1997) discusses the diversity found in these play forms from the activities themselves, player types and cultural attitudes. He affirms that these play forms do not suffice in defining play. He describes Western society's ambiguous way of thinking about play, where play is largely thought of as something useful for children to do, with the adage "learn through play", yet play is time wasting or a frivolous activity for adults. He questions how "such ecstatic adult play experiences, which preoccupy so much emotional time, are only diversions?" (Sutton-Smith, 1997, p.7). In order to address some of this ambiguity Sutton-Smith (1997) proposes seven rhetorics of play, where rhetoric is a descriptive term that describes a way of thinking about a ludic activity:

1. Progress – contemporary origin - to learn through play, predominately child or animal play, more recently serious games;
2. Fate – ancient origin – chance, gambling;
3. Power - ancient origin – sport, contests, competition;
4. Identity - ancient origin – community;
5. Imaginary - contemporary origin – art , literature, role-playing, fantasy;
6. Self - contemporary origin – solitary, hobbies, extreme sports intrinsic experiences;

7. Frivolity - ancient origin – view of play as a useless activity, jokes, pranks, comedy.

These rhetorics define play not as the activity itself, but as how the player engages with the activity. They classify player attitude and cultural attitude toward a game rather than the game itself.

A game has a common goal which has no bearing on anything that is outside of the game (Salen and Zimmerman, 2004). A game must exist within its own boundary; a frame where you enter or leave the game and this boundary is created by the decision to play. Huizinga (1949) conceives this boundary as the *magic circle* of play. To play within the magic circle helps the player to block out reality. They can become unaware of any physical boundary of the game, for example, the interface (Federoff, 2002). Flow can only happen within the bounds of the magic circle. Caillois (1961) emphasises that once reality is brought in to a game, play ceases. A game has to exist outside of reality in order for a player to immerse themselves in the activity of play or achieve a state of flow. He warns of profiteering from games; evidence of this can be seen in online games today where advertisements on so-called freemium games continuously interrupt gameplay. Metagaming is a relatively new paradigm that indicates how play is evolving with the increasing market for online and video games. The magic circle of a game is being increasingly invaded by the outside world producing peripheral stimuli (Salen and Zimmerman, 2004; Carter et al., 2012). However, little is understood about the effect this has on player enjoyment and if it challenges existing theories of play.

2.4 Defining 'Casual' in Game Design

'Casual' defines the mechanics and aesthetics of a game as much as the player behaviour. Casual refers to the characteristics of the game, the method of playing, the situation in which the game is played and the device it is played on (Kallio *et al.*, 2011). Whilst there are casual games, there are also casual players (Kuittinen *et al.*, 2007; Juul, 2009; Bateman *et al.*, 2011). The games world subscribes to two stereotypes of game: *hardcore* and *casual*. Hardcore games require huge time investments, absorbing the player into a virtual world. Gameplay consists of complex strategy and controls must be learned. The stereotypical hardcore game player is a socially awkward young male aged 16-24 (Juul, 2009). In contrast casual games can be played in short bursts of time, they are easy to learn, simple to play and offer quick rewards (Kuittinen *et al.*, 2007). The stereotypical casual game player is a woman over 30 who is unwilling to invest effort in learning a game and can only afford to play in short bursts (Juul, 2009; Kuittinen *et al.*, 2007). These stereotypes describe two types of player however, the reality is much broader. A well-made casual game could be played by people of all ages, genders and levels of gaming experience. More people play casual games than hardcore games (Fortugno, 2008; Bateman *et al.*, 2011; Kallio *et al.*, 2011). Kuittinen *et al.* (2007) defines casual games as 'games for all' as they appeal to a broad audience and span genres.

The key to designing casual games is defined by Nolan Bushnell⁴ "*All the best games are easy to learn and difficult to master. They should reward the first quarter and the hundredth.*" The phrase "easy to learn, difficult to master" is so widely used by games

⁴ <http://nolanbushnell.com/bio/>

researchers and writers that it has been named Bushnell's Law. Fortugno (2008) offers a description of the key components of a casual game. The mechanics should be intuitive and the aesthetics should provide clear interfaces. A casual game must consider the least experienced player first, allowing players to develop skills as they progress. There is no place for frustration in casual games, emphasis should be on achievement not struggle; however, the game should advance in difficulty to appeal to more skilful players, allowing them to expand their repertoire of skills (Juul, 2009). A casual game is played for stress relief, to keep the mind sharp, to kill time or as a distraction (Kuittinen et al., 2007; Kallio et al., 2011). Juul (2009) propose that games can be a creative space where the player creates and performs rather than competing. He introduces the idea of games without rules such as The Sims series or Minecraft which are flexible and accommodate many player types. Juul (2009) describe five elements of casual game:

1. Fiction – refers to the story, it describes what the game is about. This is not narrative that runs through the game as is present in many hardcore games but the graphics, the artwork, the packaging, reviews and adverts. The loose story line that sits on top of the mechanics. The story dictates the user's perception of the game and helps them decide whether or not to play.
2. Usability – describes how easy the game is to use not in terms of gameplay but as a system. Casual games need an intuitive interface and controls, they must be easy to learn, player errors should not come from bad usability and players should not be penalised for system errors.
3. Interruptability – is the ability to pace the gameplay, pausing when necessary or leaving the game without being heavily penalised. The user needs to control the time they invest in a game rather than the game demanding certain amounts of time.

4. Difficulty – is the balance between perceived skills and perceived challenge (Csikszentmihalyi, 2000). Casual games need to be easy to learn yet difficult to master. A good casual game will offer easy challenges for players with low difficulty tolerance and accommodate those players as they improve. It will also offer challenges for more advanced players who have the desire to master the game. Most casual game players have a low tolerance for frustration; they lose interest when faced with too difficult a challenge (Bateman *et al.*, 2011).
5. Juiciness – describes the use of sound, animation, or physical feedback such as vibration to support players' feelings of progression and offer rewards. A juicy interface provides positive feedback for actions.

Whilst the games can be played in short bursts there is evidence that players will invest large amounts of time playing casual games, akin to hardcore gamers (Juul, 2009; Kuittinen *et al.*, 2007; Nielsen, 2009; Kallio *et al.*, 2011). As there are two types of game, there are also two types of players: hardcore and casual. Hardcore players invest large amounts of time trying to master the game. They like difficulty and enjoy negative fiction such as fantasy, war or role-playing (Juul, 2009). Casual players play when they have the time, but have varied time investment, they are flexible. They have a low tolerance of difficulty and demand excellent usability. They enjoy positive fiction such as cute cartoons of everyday situations (Juul, 2009), and lose interest quickly and move on to new games (Kuittinen *et al.*, 2007). Kuittinen *et al.* (2007) make a distinction between a casual player who plays games in a casual way and a casual game player who plays casual games. They suggest six types of game player: *power*, *social*, *leisure*, *incidental*, *dormant* and *occasional*. The player types were derived by conducting content analysis of the views of game professionals, game journalists, industry white papers and surveys. Fortugno (2008) suggests that casual game players came to games as internet users and therefore usability is more important in

a casual game than in a hardcore game. Whereas hardcore players will see a usability fault as an obstacle to overcome and part of the game, it could deter a casual player who values ease of use as a key motivation to play.

2.4.1 Designing for Player Enjoyment – Player Type

To design for player enjoyment it is imperative that an understanding of player types is gained Dixon (2011). Cowley *et al.* (2008) define three types of player: hardcore, casual and combined. These three player types can be split into 4 types as defined by Bateman and Boon (2005): *Conqueror*, *Manager*, *Wanderer*, and *Participant*. A Conqueror is a competitive goal-orientated player; they are dominant in the game and want to win at all costs. A Manager is a logistical player; they are process-orientated and want to develop mastery of the game. A Wanderer is a curious player; they are less challenge-orientated and desire new and fun experiences. A Participant enjoys social interactions as part of play. Wanderers and Participants are more likely to play casual games or play games in a casual way, whereas Conquerors and Managers are more likely to play hardcore games or play casual games in a hardcore way (Kuittinen *et al.*, 2007; Bateman *et al.*, 2011). These player types are developed from the player types defined by Bartle (1996) of *Achiever* (Manager), *Explorer* (Wanderer), *Socialiser* (Participant) and *Killer* (Conqueror), using Myers-Briggs personality theory. Bartle's player types were defined for designers of MUD (Multi-User Dungeon) games from which MMORPGs (Massively Multiplayer Online Role-Playing Games) emerged to model what players find fun in virtual world play. They do not categorise what players do, but rather why they do it. Yee (2006) argues that Bartle (1996)'s player roles, whilst widely accepted, have never been empirically tested and the four player types may not be independent. Yee (2006) also suggests that it may be possible for a player to fit more than one player role. Through a factor

analysis of user studies, Yee (2006) exposed ten components of player behaviour; further statistical analysis revealed that these ten components could be grouped into three main components: *Achievement* (progress, rules and competition), *Social* (collaboration and support) and *Immersion* (discovery, role-playing, customisation and escapism). Bartle (2009) defends his theory of player types insisting that they were developed for designers to give them a shared vocabulary to describe different players in virtual worlds; they were never intended as a psychological model of player behaviour. He recognises the limitations of the model in that it doesn't account for immersion or how players can change from one type to another. He asserts that each player type is independent and they co-exist in balance. Changing the balance can deter players rather than encouraging more of a specific type. Tuunanen and Hamari (2012) present a taxonomy of the most popular player types derived through a segmentation process and meta analysis of related literature. These are (in order of agreement) *Achievement*, *Sociability*, *Exploration*, *Immersion*, *Skill*, *Killer*, and *In-Game Demographics*, highlighting the influence of Bartle (1996).

The majority of research into player types has been limited to modelling player behaviour in MMORGS (Kallio et al., 2011; Tuunanen and Hamari, 2012). Bartle (2009) calls for his theory to be developed further to model player types for all digital games. In a move away from Bartle (1996), Kallio et al. (2011) present a contextual model containing heuristics as a tool to aid game design rather than a classification of player types. They develop ideas of player type to account for the type of game: *Social*, *Casual* and *Committed* (hardcore); where the game is played: *Game* (the device, access, location); how the game is played: *Intensity* (session length, concentration, regularity) and with whom: *sociability* (physical or virtual space and outside of the game space). The model highlights nine reasons to play each type of game:

Social – gaming with kids, gaming with mates, gaming for company.

Casual – killing time, filling gaps, relaxing.

Committed – having fun, entertainment, immersion.

‘Casual’ defines casual game players and players who play hardcore games in a casual way. A model of gamer mentalities was created by Kallio *et al.* (2011) through a triangulation of material collated during three qualitative studies and one quantitative study. The model offers a set of heuristics for understanding why people choose to play games and how they play. The authors found no typical behaviour in their casual player type: game, sociability and intensity varied considerably. Players killing time mostly play free to download games on pc, laptop or tablet phone. They have low sociability, preferring solo games and their intensity varied with some playing for a few minutes and some playing for hours, daily or sporadically. Levels of concentration also varied, with some concentrating fully on the game and others playing whilst undertaking other activities. Players filling gaps showed more consistency, playing with short intensity and with varied concentration whilst waiting for something, taking a break or undertaking a journey for example. Players relaxing are more likely to play regularly for longer periods of time. Familiarity in a game is of most importance to this player type; they are the most likely to play with greater intensity, playing when they have time rather than making time to play. Bateman *et al.* (2011) extend their Bateman and Boon (2005) player typology by assigning psychometric models to player types to deepen the understanding of why people play, resulting in four revised player types: *Logistical*, *Tactical*, *Strategic* and *Diplomatic*. The authors claim that player type theories are inadequate and suggest a move toward modelling player traits. They suggest five traits: *openness to imagination*, *preference for anger vs. avoidance of frustration*, *degree of tolerance for real time play*, *group play vs. solo play* and *persistence*. No formal model has

been published to date. Fullerton (2008) describes ten player types: *The Competitor*, *The Explorer*, *The Collector*, *The Achiever*, *The Joker*, *The Artist*, *The Director*, *The Story Teller*, *The Performer* and *The Craftsman*. Derived from the theories of Caillois (1961), Sutton-Smith (1997) and Bartle (1996), these player types are more thorough as they also model theories of play. In particular, *The Collector* has not been defined in other literature. These players find enjoyment in acquiring items, trophies, badges and knowledge; they create sets and organise.

Kallio *et al.* (2011), Tuunanen and Hamari (2012), Yee (2006) and Dixon (2011) claim that no one player can be categorised into a single player type. They are, however, useful as a design tool in structuring Dynamics, Mechanics and Aesthetics to appeal to a broad range of players (Bateman *et al.*, 2011; Bartle, 2009; Dixon, 2011).

2.4.2 Designing for Player Enjoyment - Fun Factors

Knowledge of player types provides the insight to develop a game experience that will appeal to many users. To understand what players enjoy about a game is to understand what makes a game fun. Febretti and Garzotto (2009) argue that fun is the fundamental factor that motivates users to continue to play a game over time. Fun is subjective and will vary from person to person; however, player types offer components that can be used to design experience. Methods to design for fun have been researched by Lazzaro (2008), Juul (2009) and Hunicke *et al.* (2004). Hunicke *et al.* (2004) describe three components of games: Rules, System and Fun, and claims that their design counterparts are Mechanics, Dynamics and Aesthetics. Mechanics describes the algorithms in the game that create the rules and the gameplay. Dynamics describes player interaction and feedback (inputs and outputs). Dynamics

create Aesthetics; Aesthetics describe emotional responses and whether the player is enjoying the game. The authors define aesthetics using eight components or a 'taxonomy of fun':

1. Sensation – game as sense pleasure;
2. Fantasy – game as make believe;
3. Narrative – game as drama;
4. Challenge – game as obstacle course;
5. Fellowship – game as social framework;
6. Discovery – game as uncharted territory;
7. Expression – game as self-discovery;
8. Submission – game as pastime.

(Hunicke et al., 2004, p.2)

The influence of player types is evident in these components. Multiple player types can find the same aesthetics fun whilst having very different goals. Hunicke *et al.* (2004) argue that not all components need to be applied for a player to enjoy a game and that games will emphasise some components over others.

Lazzaro (2008) developed the concept of *Four Fun Keys*:

1. Hard Fun – challenge and mastery;
2. Easy Fun – imagination, exploration and role-play;
3. Serious fun – doing real work or immersion/flow;
4. People fun – social interaction.

Fun Keys categorise game elements and distinguish how a person plays a game. Each Fun Key can be found in either hardcore or casual games. The most successful games will have at least three of these Fun Keys. Fun is the series of choices and

feedback that lead to the flow state. Lazzaro (2008) claims that a player will feel many emotions in the 'flow zone' defined by Csikszentmihalyi (2000) as the range between boredom and anxiety. Game designers can design Mechanics, Dynamics and Aesthetics that can induce certain emotions in a player type and improve the probability of flow experience. Each Fun Key describes a series of choices and feedback that stimulate players creating a certain set of emotions that contribute to a sense of flow. To experience Hard Fun designers must develop mechanics that suggest multiple strategies and create numerous obstacles and multiple goals. The player needs to be stretched harder to progress through challenges and master the game. Flow emerges from feelings of frustration and relief, finding success only after being pushed to the point of quitting. A player looking for Hard Fun will play any game in a hardcore way. Easy Fun inspires imagination and is induced by curiosity. It is not initiated by challenge but by the enjoyment of interaction. Easy Fun is creative and explorative; it is an alternative to Hard Fun and can often be used as a method of capturing player attention between challenges. Easy Fun could be found in creative pursuits such as designing scenery, avatars and characters, experimenting with the controls, or in trying to break the game. Both Hard Fun and Easy Fun are developed from flow theory of an optimum state between boredom and anxiety. Hard Fun balances frustration and boredom; if a challenge is too difficult, a player will become too anxious and stop playing, if it's too easy the player will feel bored and stop playing. Easy Fun balances disbelief and disinterest; if the game experience is too novel a player will not know how to play, creating anxiety, but if it is too predictable then there will be nothing for the player to explore, causing boredom.

Table 2-1 shows player types mapped to the Fun Keys. Player types cannot be mapped to Serious Fun; immersion is a state that all player types can reach rather

than being a type in itself, and as the majority of research into player behaviour has focussed on hardcore, fantasy, role-playing games, they did not model real work. In the case of Yee (2006), the player type *Immersion* is used to describe a number of behaviours that are associated with the narrative, fantasy and role playing elements of MMORPG's, rather than the immersive state that creates flow. Serious Fun represents the perceived purpose of the game and the desired outcome at the end of a series of tasks. Serious Fun is therefore the ultimate goal for each player type. To experience Serious Fun, players need to be engaged with the game emotionally and mentally, they play with purpose or use games as therapy. Players will have a reason to create something of value outside of the game itself, be that to relax or kill time (Kallio et al., 2011), complete a useful task as in the case of GWAPs, or generate reputation or bragging rights. Stimulation comes from the individual's thoughts about the game rather than curiosity or imagination and it is possible to engage in the game without challenge. The optimum outcome of Serious Fun is how a player values the game experience.

Lazzaro (2008) suggests that People Fun is the least needed of the four Fun Keys, but that games without People Fun need to be exceptionally strong in other Keys. Bateman *et al.* (2011) corroborate this statement empirically, finding that 40.6% of their user study respondents prefer to play alone. In the same study, they report that the majority of the respondents do not get enjoyment from Hard Fun, detailing that they avoid games which create anger or too much frustration. Interestingly, the majority of players who stated that they enjoy Hard Fun also enjoy playing multiplayer games. Providing social interaction into usually solitary gameplay can enhance the game experience, e.g. Xbox live, in game chat, integration with social networks such as Facebook, and online communities. The game experience can be

personalised to improve performance. People Fun provides a relationship with other players and creates opportunities to compete or to collaborate.

Table 2-1 Mapping fun factors to player types.

Fun Keys	Player Type			
	Bartle (1996)	Bateman and Boon (2005)	Yee (2006)	Bateman <i>et al.</i> (2011)
Hard Fun	<i>Killer</i> <i>Achiever</i>	<i>Conqueror</i> <i>Manager</i>	<i>Achievement</i> <i>Immersion</i>	<i>Logistical</i> <i>Tactical</i> <i>Strategic</i>
Easy Fun	<i>Explorer</i>	<i>Wanderer</i>	<i>Immersion</i>	<i>Diplomatic</i>
Serious Fun				<i>Logistical</i> <i>Strategic</i>
People Fun	<i>Socialiser</i> <i>Killer</i>	<i>Socialiser</i> <i>Conqueror</i>	<i>Social</i>	<i>Diplomatic</i>

2.4.3 Measuring Enjoyment

Caillois (1961) offers a classification of games, defining four game types: *Agon*-competition; *Alea*-chance; *Mimicry*-role playing and *ilinx*-flow. These four types operate along a scale. At one extreme is *Ludus*, with rules and structured activities. At the other is *Paidia*, with spontaneous, unstructured activities. Salen and Zimmerman (2004) suggest that the Caillois (1961) model of classifying types of game can be used to identify what play experiences your game is or is not providing. Isbister (2010) advocates that research should now concentrate on using available models to encourage user behaviour through design rather than creating more taxonomies of user behaviour. User behaviour models offer an indication of what aspects of games

different types of players will find enjoyable. Salen and Zimmerman (2004) surmise that flow is just one of many tools available to describe player enjoyment. They highlight how flow might not be the best suited measure of enjoyment as it is not unique to games, is more about the player than the game and is not a universal phenomenon. However, theory of flow is the foremost theory applied to games design research and is seen as the benchmark in attempting to understand and begin to measure user enjoyment of games (Lazzaro, 2008; Cowley et al.,2008; Shin and Shin, 2011; Jegers, 2007; Voiskounsky et al.,2004; Chen, 2007; Chiang et al., 2008; Hsu and Lu, 2004; Sweetser and Wyeth, 2005). Enjoyment is difficult to measure (Carroll and Thomas, 1988). Therefore, much of the research in this area has remained theoretical, mapping elements of flow to existing concepts of game design with little movement to create tested frameworks that enable designers to attempt to predict whether a game will be enjoyable, and so having the potential to be popular or to provide methods to measure enjoyment.

Hsu and Lu (2004) discuss an extended TAM (Technology Acceptance Model) which combines theories of perceived usefulness and perceived ease of use with concepts of flow to model why people play online games. User studies revealed that Perceived Usefulness (PU) does not motivate users to play games but it directly affects their attitude toward playing the game. PU has a significant link with attitude, but not with a player's intention to play a game; flow was found to have a significant link to intention, but not to attitude. Similarly, Shin and Shin (2011) extend a TAM to predict social game acceptance and model user behaviour of social network games. They model some different factors to Hsu and Lu (2004): PU, Attitude, Intention and Flow are maintained, but Perceived Ease of Use is replaced with Perceived Enjoyment (PE), Perceived Playfulness (PP) and Perceived Security (PS). The factors are

categorised as either Positive or Inhibiting, PS and PU being inhibiting, PE and PP being positive. User studies revealed that both PS and PP had a significant effect on attitude, with PP having a stronger effect than PS. PE was found to have a significant effect on intention. Improved perception of flow was found to increase PP. In contrast to Hsu and Lu (2004), Shin and Shin (2011) found that PU has no significant effect on player attitude but established a significant link to intention. This discrepancy is indicative of the subjective nature of user behaviour, especially when dealing with a small group of users. Shin and Shin (2011) describe the limitations of their own research and explain that their evaluations identified important paths to flow that the model had missed.

Sweetser and Wyeth (2005) map the theory of flow to existing usability heuristics for game design and develop a model that can be used to evaluate player enjoyment. The Game Flow model offers game designers a new way to make design decisions that could increase player enjoyment or discover aspects of the game that could be improved. The authors summarise how the elements of flow occur during gameplay to manifest in complete engagement with the game. The player's attention must be kept through high work-load. Tasks should be sufficiently challenging to match a player's skills and tasks must have clear goals so that the player knows what they need to complete. The player must receive feedback on their progress toward completing a task. They will then feel a sense of control over their actions, ultimately resulting in a feeling of complete absorption in the game, an altered sense of time, a loss of concern for self and a loss of awareness. Jegers (2007) extends the game flow model to apply it to pervasive games. The Sweetser and Wyeth (2005) model is detailed fully in Chapter 6 along with discussion of other methods of usability and playability evaluation.

2.5 Gamification

Gamification, whilst rooted in HCI and pervasive computing theories (Malone, 1982; Carroll and Thomas, 1988; McGonigal, 2011; Hamari and Koivisto, 2013; Deterding et al., 2011a) and predominately applied to education and learning (Hamari et al., 2014), is in its infancy and as such, few theoretical frameworks or empirical evaluations exist (Hamari et al., 2014; Mekler et al., 2013). The premise of gamification was conceptualised in the 1980s first by Malone (1982) and then by Carroll and Thomas (1988). Despite a plea by Carroll and Thomas (1988) for more HCI and usability research to be conducted on applying theory of fun to existing methods and beliefs on system design, progress in this area has been slow. Playability and user experience are relatively recent and industry centred developments. No area captured the essence of the Malone (1982) theory until the emergence of GWAP and gamification. Malone (1982) describes systems as being *tools* or *toys*, distinguishing between toys (games) used for their own sake with no external goal and tools as systems used as a means to achieve external goals. He hypothesises that if the external goal of using a system is not highly motivating (i.e., routine and boring), toy-like features can be useful in making the activity enjoyable. A framework of heuristics is proposed to analyze the appeal of games that can be used as a checklist to design enjoyable interfaces. The framework consists of three categories: *Challenge*, *Fantasy* and *Curiosity*. *Challenge* refers to goals, uncertain outcomes and variable difficulty; *Fantasy* refers to the overall look and feel of the system and metaphors induced from the aesthetics, *Curiosity* relates to novelty of the interaction and juiciness of feedback.

A view is beginning to emerge in the literature that GWAP is a form of gamification (Liu et al., 2011; Hamari et al., 2014; Venhuizen et al., 2013; Gomes et al., 2013b). The

term is also appearing more in titles of papers where the content refers to other topics, predominantly GWAP and serious games (Hamari *et al.*, 2014). Mekler et al. (2013) notes that a Google Scholar search for 'gamification' netted 1,780 publications as of December 19th 2012, with 1,180 published only in 2012; 4,780 results were returned for a search for 'gamification' on 5th February 2014: 3,000 more papers in a little over 12 months. The term 'gamification' was coined in 2008, but it was not until 2010 that it began a transition from digital media industry buzz word to an area of academic interest (Deterding et al., 2011b; Huotari and Hamari, 2012). The research has concentrated on defining gamification (Huotari and Hamari, 2012; Hamari and Koivisto, 2013) and most notably (Deterding et al., 2011a; Deterding et al., 2011b). Nicholson (2012) provides a theoretical framework for *meaningful gamification*. Hamari and Koivisto (2013) present findings of an empirical evaluation on a gamified system.

Deterding *et al.* (2011b) provide the most commonly cited reference of gamification as "The use of game design elements in non-game contexts". Gamification is a method of improving user experience and creating joy of use. Huotari and Hamari (2012) argue that relying on game elements to refer to gamification is wrong as there is no clearly defined set of game elements. There are no aspects of games that automatically create gameful experiences: experience is player dependant. (Huotari and Hamari, 2012, p.19) suggest a definition of gamification as "a process of enhancing a service with affordances for gameful experience in order to support a user's overall value creation"; in other words, designing more playful interfaces to improve user productivity and output. Huotari and Hamari (2012) argue that creating a gameful experience takes more than a scoring system, but do not offer a definition of gameful experience. Gameful experience is defined by Deterding *et al.*

(2011a) as flow. As previously discussed, flow is not reserved for games and game is not a pre-requisite for flow, but flow can occur through the play of a game. Flow comes through play, not use of game mechanics. Gamification relates to game, not play. Deterding *et al.* (2011a) defines gamefulness as qualities of gaming and playfulness as behavioural qualities of play. Relating to Caillois (1961) playfulness is *Paida*, Easy Fun (Lazzaro, 2008) and gamefulness is *Ludus*, Hard Fun (Lazzaro, 2008). Deterding *et al.* (2011a) assert that gamification of a system promotes gamefulness, not playfulness, potentially alienating Explorers and Wanderers seeking novelty and Easy Fun, and the majority of players prefer Easy Fun (Bateman et al., 2011). Users seeking Hard Fun crave competition, strategy, mastery; Hard Fun attracts Achievers, Killers and Conquerors. What quality of rule based, goal orientated mechanics can realistically be applied through gamification when the emphasis has to be on the underlying utility of the system? Chorney (2013) and (Robertson 2010, in Nicholson 2012) advocate that points, levels and leaderboards are not what make a game and that it is insulting to the game genre to assume game enjoyment is so simple to replicate. Gamification is nothing more than adding a scoring mechanism to non-game activities (Robertson 2010, in Nicholson 2012).

Deterding *et al.* (2011a) stress that the boundary between game and artefact is blurry: when is it a game or a gamified application? The difference is in the user interaction: a game is played, an application is used. A gamified system provides an unstable experience between playful, gameful and general use (Deterding et al., 2011a). Game design elements are building blocks toward an experience not required components. Gamified systems are built with the intention to include elements from games rather than a full game. A user can 'play' the system or use it without engaging in game elements. The effectiveness of game elements depends on the service in which they

are used. The addition of game elements cannot transform a system into a game but they can bring about a gameful experience (Hamari, 2013). Deterding *et al.* (2011a) highlight the problem of how to identify game elements and Deterding *et al.* (2011b) emphasise that gamification uses elements of games, not elements of play. They propose five levels of game design elements at varying degrees of abstraction, listed by most concrete first:

1. Interface Design Patterns – badge, leaderboard, level;
2. Game Design Mechanics – time constraint, limited resources, turns;
3. Game Design Principals and Heuristics – enduring play, clear goals, variety of game styles;
4. Game Models – MDA (Hunicke et al., 2004), Game Type, Challenge, Fantasy, Curiosity (Malone, 1982), Lupus and Paidia (Caillois, 1961);
5. Game Design Models – play testing, play-centric design, value conscious design.

Using these levels as guidelines, designers can select appropriate game elements for their system. To apply a scoring mechanism to a system would entail only applying one level and at a simple, concrete level of abstraction. The success of a gamification project may lie in the level of abstraction of the game elements applied. By applying only one game element at the lowest level of abstraction, badges, Hamari (2013) found no evidence of competition or increased usage frequency or quality of social interaction. They suggest a low goal commitment; it is obvious that collecting badges was not enough of a goal to users. This is an example of how gamification can be used badly and could actually deter use (Nicholson, 2012).

Hamari and Koivisto (2013) define gamification as follows:

- Affording and creating experiences reminiscent of games involving flow, mastery and autonomy;
- Attempting to affect motivations rather than attitude or behaviour;
- Adding gamefulness to existing systems rather than building an entirely new game.

Their definition focuses more on using gamification to affect how the user interacts with and perceives the system rather than on affecting the interface with which the user interacts. Whereas Deterding *et al.* (2011a) focus on gamefulness, Hamari and Koivisto (2013) focus on playfulness, whilst referring to it as gamefulness, emphasising the contradictory nature of the literature and lack of clear understanding of what gamification achieves. According to Hamari and Koivisto (2013), a user's motivation to use a gamification system is reliant on social influence. This is similar to findings by Hsu and Lu (2004) and Shin and Shin (2011), Hamari and Koivisto (2013), observed in empirical evaluations that perceived use and perceived socialness of the system motivates use. Perceptions of continued use were affected positively by the presence of social feedback for actions (e.g., likes and comments). The authors suggest gamification has more chance of success when a community of users committed to the goals of the system already exists.

Users are more likely to engage in behaviour they perceive others are engaged in (Hamari, 2013). Groh (2012) suggests that for gamification to work, the user needs to be introduced to a meaningful community with the same interests and that this can be achieved through aesthetics. He warns, however, that over-associating a system with a specific social context meaning could alienate potential users outside of the special interest group. Hamari and Koivisto (2013) reiterate the Deterding *et al.* (2011a) approach that gamification is applied to an existing system and that it creates

a system layer of game mechanics that makes a system more engaging. Groh (2012) notes the distinction between game and gamified system in that if the gamification layer is removed from a gamified system, there is still meaningful content. Nicholson (2012) argues that to develop, gamification needs to focus less on mimicking a game and more on how each game element will benefit the user experience. Using external rewards to control behaviour creates a negative feeling and the focus is on system output rather than user experience. Gamification can be harmful to system use if not executed appropriately. Nicholson (2012) propose a theoretical framework of meaningful gamification where the gamification is user-centred not system centred. In contrast to Deterding *et al.* (2011a), and similar to Hamari and Koivisto (2013) he promotes playful design not gameful design; game elements need to be deeply integrated in the system. Users are then less aware of the game elements and are therefore less likely to feel controlled or exploited by them.

There is no monetary incentive to use a gamified system; users are rewarded by entertainment. Users have less incentive to enter low quality or malicious data if they freely choose to engage in the activity rather than being controlled by the extrinsic motivation of financial reward (Liu et al., 2011). Reimer (2011) uses Huizinga's (1949) theories of play to re-evaluate gamification and discusses how it can be developed from its current position. He argues that the presence of game elements in a system is not enough to motivate a person to play. Play cannot be ordered, although a space can be created where play is encouraged, but not essential. If the game elements are removed or not engaged with, a user can still use the system and gain enjoyment from it. Reimer (2011) suggests that gamification needs to be rethought as a method to develop a 'magic circle' as Huizinga (1949) theorised and not just a process of adding game elements. He claims that leaderboards hinder

motivation and instead focus should be on mechanics that provide constant and clear feedback. He advocates that gamification should concentrate not on competition but on encouragement that the task is being completed well. Mechanics and Aesthetics should be used to reveal a magic circle; the magic circle will entice users in.

In its current state gamification rewards use, it is a form of extrinsic motivation. This can deter gameful experience and does not incite play (Cherry, 2012). Play cannot be controlled, it must be a freely chosen activity and users are intrinsically motivated to play. Participation in a gamified system is extrinsically motivated Bouca (2012). Groh (2012) and Aparicio *et al.* (2012) apply Ryan and Deci's (2000) Self Determination Theory to Deterding *et al.*'s (2011b) definition to propose methods for gamification that encourages intrinsically motivated users. Carroll and Thomas (1988) warn that sources of intrinsic motivation in games might not have the same effectiveness when implemented in a non-game context. Groh (2012) proposes three core components of a gamified system: *Relatedness*, *Competence* and *Autonomy*. Relatedness describes a user's perception of the system, or perceived usefulness and perceived socialness; how they perceive goals, community, and the fiction. Competence describes perceived ease of use and also a user's perception of enjoyment. This encompasses game design elements at various levels of abstraction. Finally, Autonomy warns of devaluing the original system with gamification. A user must feel that engaging with the gamification layer is voluntary. Aparicio *et al.* (2012) develop the idea further by describing a method for an effective process of gamification. Three tasks are described: first, identify the main objective - the goal, the purpose, competence; second, identify the transversal objective - how interesting users will find it, relatedness; third, selection of game mechanics. Game mechanics are grouped by the intrinsic motivation it is hoped they will promote:

- Relatedness - groups, messages, blogs, connection to social networks, chat.
- Competence – positive feedback, optimal challenge, progressive information, intuitive controls, points, levels, leaderboards.
- Autonomy - profiles, avatars, macros, configurable interface, alternative activities, privacy control, notification control.

The methods for gamification, or game design in a GWAP, are only one aspect of the design process. The overriding facet is the purpose of the system - why are users being encouraged to use it? The purpose creates the goal for the player and the output for the system. In video tagging games, users could potentially be as motivated by tagging videos as they are by the game elements. Successful gamification and video tagging game design is an engineered balance between user enjoyment, accurate input and output that is fit for purpose.

2.6 Tagging

Tagging emerged as one of the fundamental characteristics of the social web. Tagging is the process of assigning free form labels to web resources. It is a sense making activity where meaning emerges through categorisation and labelling of online content (Golder and Huberman, 2006). Shirky (2005) describes tagging as free form labelling with categorical constraint. It allows for resources to be placed in multiple categories which reduces the cognitive cost to the user of finding the one perfect category (Mathes, 2004). There is “no hierarchy, no directly specified term relationships, and no pre-determined set of groups or labels used to organize user’s tags.” (Lin et al, 2006, p.2). Tagging is a low cognitive cost activity, (Furnas et al., 2006; Mathes, 2004; Sinha, 2005) although Sen *et al.* (2006) found that 68% of users

found it difficult to think of a tag to describe a movie and 51% of user tags are reused. Chi and Mytkowicz (2007) identify 'lazy taggers' who would rather use a tag suggestion than think of one of their own. Tags reflect the vocabulary of users (Mathes, 2004). They provide contextual and dynamic information that cannot be derived from the resource itself (Lee and Hwang, 2008). Tags can inform a system about user characteristics and attitudes that can be utilised for personalised search or recommendations (van Velsen and Melenhorst, 2009). Tagging can be a gameful experience, users get immediate feedback on their input, there is community (Furnas *et al.*, 2006) and users have clear goals (Melenhorst and van Velsen, 2010). Wu *et al.* (2006) advocate using tags to improve metadata as they are user generated and so match the natural language people use in search (Tjondronegoro and Spink, 2008). Lee and Hwang (2008) describe tags as additional metadata that can be used as rating or reviews of resources.

The type of tagging system affects how users will tag. Marlow *et al.* (2006) and Sen *et al.* (2006) describe tagging system characteristics (summarised in Table 2). Tagging is a tripartite network where tag, user and resource are inter-related (Cattuto *et al.*, 2007; Lambiotte and Ausloos, 2006; Marlow *et al.*, 2006; Halpin *et al.*, 2007; Kern *et al.*, 2008). Voss (2007) suggests that the tripartite graph is too simple as users, resources and tags could be linked independently of the tagging system. This is becoming more apparent with developments in linked data and RDF triples (Passant, 2007; Hildebrand and Ossenbruggen, 2012). Tag sets in relation to the resource are called a folksonomy, a term coined by Vander Wal (2007) to mean a taxonomy created by folk. Depending on the tagging system, a folksonomy can be *Broad* or *Narrow*. Broad folksonomies are created by collaborative or social tagging systems where many users describe many resources. Tags can be assigned by the content

creator and the content consumer (Wu et al., 2006). Narrow folksonomies are created by owner tagging systems where one user tags resources they upload to the system. Tag sets in relation to the user are called a personomy, the collection of tags assigned by an individual user. Whilst folksonomy is useful for categorising resources, personomy is useful for search.

Table 2-2 Summary of tagging system characteristics.

<p>Tagging Rights (Marlow et al., 2006)</p>	<p>The tagger's relationship with the resource:</p> <p>Owner tagging – users only tag resources they upload to the system, many-one relationship.</p> <p>Collaborative/social tagging - users assign tags to any resource in the system, many-many relationship.</p>
<p>Tagging Support (Marlow et al., 2006)</p> <p>Tag Sharing (Sen et al., 2006)</p>	<p>Whether existing tags are visible to taggers and whether suggestions for tags are given.</p> <p>Blind, Suggested or Viewable. Only showing the most frequently added tags. Tag Clouds.</p>
<p>Aggregation (Marlow et al., 2006)</p> <p>Other Dimensions (Sen et al., 2006)</p>	<p>Whether the system allows tag repetition.</p> <p>Whether tag frequencies can be calculated.</p> <p>How the system allows users to enter tags.</p> <p>Single or Multi-word, compound, punctuation, white space.</p>

Type of object (Marlow et al., 2006)	Image, video, web page, piece of music.
Source of Material (Marlow et al., 2006)Item Ownership (Sen et al., 2006)	User upload (YouTube), system provided (GWAP) or another user's.
Resource Connectivity (Marlow et al., 2006)	How resources are linked - internal or external.
Social Connectivity (Marlow et al., 2006)	How users are linked - internal or external.

Voss (2007) define four user roles in a tagging system: *Resource Author* – the creator; *Resource Collector* – adds the resource to the tagging system; *Indexer or Tagger* - person that tags resources and *Searcher* – person that uses tags to find resources. Users can fulfil different roles at the same time. Furnas *et al.* (2006) question whether enough users understand the process of tagging, suggesting that only the technology savvy tag, not the masses. If the masses tag, will they find the role of tags and find usefulness in them? Lack of understanding will affect the types of tag they enter and their motivation to tag. Melenhorst *et al.* (2008) compared the tags entered by three types of tagger, BasicTagger (no suggestions), SocialTagger (suggestions from user tags) and LazyTagger (suggestions from user tags and professional annotations); LazyTaggers produced the most tags with a statistically significant difference to BasicTaggers. LazyTaggers produced more unique tags and tags considered useful for search.

Heralded as the solution to the problem of indexing the web (Shirky, 2005), tagging gained considerable academic interest, particularly in information and library science, highlighting its positives and negatives compared to traditional indexing

methods and controlled vocabulary (Macgregor and McCulloch, 2006; Kipp, 2006; Kipp, 2007b; Kipp, 2007a; Guy and Tonkin, 2006; Ding et al., 2009; Voss, 2007). The majority of research measures quality as relevance, and relevance is dependent on user agreement. Therefore the few tags that are entered most frequently are deemed to be the highest quality tags (Halpin et al., 2007; Gligorov et al., 2011; Sood et al., 2007; Lin and Aroyo, 2012). Unique tags are not useful. They are irrelevant and can be discarded in the pursuit of efficiency (Halpin et al., 2007; Lin and Aroyo, 2012). Lin *et al.* (2006) and Kipp and Campbell (2006) found a tag frequency ratio for unique tags and tags with high agreement to be 30/70, not the predicted given by the 20/80 Pareto Principle suggesting that taggers have high agreement on basic level terms. Kipp and Campbell (2006) claim that this provides evidence that people use tagging to classify resources. (Lin et al., 2006, p.6) highlight this classification practice: “users do not attempt to tag all content of a document but instead they highlight specific content or facts most interesting to them”. Perception of quality is dependent on how the tag will be used, either classification or search.

Problems identified with tag quality, namely the vocabulary problem first discussed by Furnas *et al.* (1987), are predominantly problems for traditional indexing and classification (Aurnhammer et al., 2006; Begelman et al., 2006; Voss, 2007; Guy and Tonkin, 2006; Ding et al., 2009; Körner et al., 2010). The vocabulary problem is described in Table 2-3. In search, the variety of terms creates a precision problem because the system cannot distinguish between the multiple terms, or multiple meanings, and so returns all results. A recall problem occurs if retrieval is not extended to synonyms, plurals etc. (Furnas et al., 1987; Ransom and Rafferty, 2011). For classification, the vocabulary problem creates noisy categories and a lack of

shared vocabulary and controlled vocabulary (Sood et al., 2007; Voss, 2007; Macgregor and McCulloch, 2006; Guy and Tonkin, 2006; Halpin et al., 2007).

Table 2-3 The vocabulary problem for tagging systems.

Problem	Description
<i>Polysemy</i>	<p>One word which has many meanings (e.g., Table - a piece of furniture or a way to display information; or Apple – a piece of fruit or a computer company).</p> <p>(Golder and Huberman, 2005; Wu et al., 2006)</p>
<i>Synonyms</i>	<p>Many words with the same or similar meaning (e.g., Funny: Amusing, Humorous, Comical, Hilarious, Hysterical).</p> <p>(Mathes, 2004; Golder and Huberman, 2005).</p>
<i>Ambiguity</i>	<p>Users can enter different tags to describe the same object (e.g., New York, Big Apple, NYC, USA, City, US city).</p> <p>Mathes (2004)</p>
<i>Spaces, multiple words and compound tags</i>	<p>Tagging input varies between systems, some sites allow multiword tags, some systems use a delimiter such as space to allow users to enter lists of tags rather than a single tag at a time, resulting in users using punctuation to enter multi-word tags or joining words together to form a compound tag. Whilst these tags are meaningful to the user, to others in the same system and can be used for communication or become shared vocabulary, they are less useful for search or categorisation (e.g., I am a sentence, I_am_a_sentence, Iamasentence).</p>

	(Mathes, 2004)
<i>Spelling errors and plurals</i>	Miss-spelled words and plurals appearing alongside singular versions of the same word. (Kipp and Campbell (2006), Golder and Huberman (2005)
<i>Lects</i>	Words that are unique to a person's geographical location (dialect), ethnicity (ethnolect) or social group (sociolect). People from different locations use different words that mean the same thing and use words that are only meaningful to people from that location. (Marlow et al., 2006; Kipp and Campbell, 2006; Golder and Huberman, 2005)

Tag quality is important in social tagging systems where only a selection of tags is shown to the user. As most users reuse tags they have seen, it is important that the viewable tags are of high quality. In GWAP, where the tag set creates additional searchable textual data, quality is not important for reuse but in reducing noise so that there are more high quality tags which describe the resource rather than a huge set of low quality tags and tag noise. An understanding of what constitutes a quality tag for tagging games and how these can be encouraged through gameplay is important. Tag quality can be inferred by two measures: implicit (user behaviour) or explicit (user rating). A high quality tag will enhance browsing or search and be a source of descriptive information (Sen et al., 2007). Sen *et al.* (2007) found that only 21% of tags on Movielens were worthy of display, similarly Ransom and Rafferty (2011) found that a large proportion of tags in Flickr were synonyms. In contrast, Trant (2009) only discarded 6.7% of tags due to the vocabulary problem. Al-Khalifa and Davis (2007) analysed a sample of del.icio.us tags for evidence of the vocabulary

problem: spelling errors consisted of 6% of all tags, compound tags 30% and abbreviations 6%. In del.icio.us, tags that are used for personal or social organisation are the most likely to show evidence of the vocabulary problem and tags that describe facts about the resource are the most reliable.

The vocabulary problem in tagging can be described as noisy tagging or meta-noise. It decreases tagging system utility and makes it harder for people to find resources. Wu *et al.* (2006), Begelman *et al.* (2006) and Brooks and Montanez (2006) advocate creating controlled vocabulary to reduce meta-noise and increase tag agreement to improve search. Shirky (2005) asserts that you can extract value from big messy datasets and that there is no need to control the vocabulary. Problems are associated with tagging as a method of classification. Merholz (2005) argues that tags do not reflect classification or categorisation but how users are describing resources. Rafferty and Hilderley (2007) propose democratic indexing as a method of adding some control to user tags without compromising their free-form nature. For this, a subject-based index is created from user interpretations of the content. It is flexible and can change over time as opposed to rigid groupings based on professional classification categories. The method does not index perceptual information or tags that are already available in textual data, but indexes instead subjective interpretations of what the resource is about. The system allows for both public and private indexes which can reduce the amount of personal tags in the public index that often contain more examples of the vocabulary problem (Sen *et al.*, 2007; Al-Khalifa and Davis, 2007).

Berendt and Hanser (2007) argue that tags are not metadata that can be used for classification, they just provide more content. There is no gold standard which all good metadata should adhere. Tags do not classify but produce more searchable content. As a method of describing resources and extra textual data to improve search, the broader the tag dataset the better (Leyssen et al., 2012; Furnas et al., 1987; Sheng et al., 2008). As folksonomies are uncontrolled; adding control to the vocabulary, some researchers believe, is key to filtering tag sets and overcoming the vocabulary problem (Macgregor and McCulloch, 2006; Voss, 2007). User tags are subjective and inconsistent, expert tags are objective and consistent, providing structure and control (Aurnhammer et al., 2006). Transforming the free form nature of tags into a shared or controlled vocabulary either by suggestions (Körner et al., 2010; Sood et al., 2007; Bar-Ilan et al., 2006; Guy and Tonkin, 2006), similarity clustering (Kipp and Campbell, 2006; Capocci and Caldarelli, 2008), co-occurrence or semantic clustering (Cattuto et al., 2007; Begelman et al., 2006; Kim, 2011; Körner et al., 2010), tag suggestion and tag recommendation systems (Wang et al., 2012; Graham and Caverlee, 2008; Chirita et al., 2007), or as Kern *et al.* (2008) recommend, extending the folksonomy with other metadata such as, title or user comments.

2.6.1 Types of Tag

Whilst tags can be entered with various synonym and spelling variances, they are also at varying levels of categorisation. At the conceptual level rather than perceptual, there are three levels of categorisation with a hierarchical relationship: Superordinate, Basic and Subordinate (Rosch, 1975; Croft and Cruse, 2004; Jaimes and Chang, 1999). Basic level theory has been used to categorise tags by numerous researchers: Hollink *et al.* (2004), Golder and Huberman (2005), Macgregor and McCulloch (2006), Rafferty and Hilderley (2007), Rorissa (2008) and Stuart (2012).

Rosch (1975) showed people use basic level vocabulary in free naming tasks and when thinking of the world around them. Users will enter more basic level tags than subordinate and superordinate is the least used level (Golder and Huberman, 2006; Stuart, 2012). Superordinate is the top level, the parent. It is not the first level a user will think of. It groups basic level tags e.g., animal, human, automobile. Superordinate is less useful than basic level as it has fewer defining attributes, it is the most general. Basic level is a child of a superordinate tag. It is the first level a user will think of and is defined by Croft and Cruse (2004) as the most inclusive level to which a clear visual image can be formed and characteristics drawn using general terms. Little knowledge or cognitive effort is required e.g., dog, man, car. Subordinate level tags are the children of basic level tags; they are more specific, providing a high resemblance to the object. More abstract than basic level tags they require knowledge about the subject they are categorising e.g., labradoodle, Justin Beiber, Ford Fiesta. The level of specificity people tag at depends on their subject knowledge (Begelman et al., 2006; Golder and Huberman, 2005). Enser (2008) found that in describing images people used more high level reasoning and subjective interpretation of the content than simply identifying objects, scenes or activities present in the image. With this in mind, basic level theory which only identifies the level at which 'objects' are identified is insufficient to categorise tags on its own. When tagging visual resources further segmentation of vocabulary level can be applied using the levels of interpretation proposed by Panofsky (1970): *pre-iconographic*, *iconographic* and *iconological*. Shatford (1986) generalised Panofsky's theory simplifying pre-iconography as generic interpretation and iconography as specific interpretation (Enser, 2008). Shatford (1986) created a method of image interpretation that concentrated on what the image was 'of' or 'about' where 'of' is objective and 'about' subjective. Basic level theory classifies objective and fairly concrete descriptions, superordinate and basic level correspond to the pre-

iconographic level of meaning which is the recognition of easily identifiable objects and factual information relating to an image. Perceptual features would be classed as pre-iconographic. Iconographic level requires more familiarity with the subject to identify objects at a specific level and therefore corresponds to subordinate level. Descriptions at an iconological level provide an interpretation of the visual content at the most abstract and subjective level. This level is not modelled by basic level theory and as such extends the level at which images are described. Iconological level refers to the subjective interpretation of what the content means, this can have varying levels of abstraction and depends on the subject matter for instance, more interpretation may be derived from a renaissance painting (as Panofsky's levels of meaning were originally intended) than a photograph of a group of friends in a park. Example tags can be found in Table 2.4 which shows how the various levels of interpretation are connected. Panofsky's levels of meaning are discussed in further detail in Section 7.1.

2-4 Example tags categorised by basic level theory and Panofsky's levels of meaning.

Objective			Subjective
Pre-iconographic		Iconographic	Iconological
Superordinate	Basic	Subordinate	
Animal	Dog	Labradoodle	Furry, shaggy, loopy, friendly, lovable, loyal, companion.
Human	Man	Justin Beiber	Downfall, hot, talented, untalented, idiot, loser, obnoxious, selfish, douchebag, cute, cool.
Vehicle	Car	Ford Fiesta	Small, trendy, powerful, hatch, reliable, economical.

Basic level theory and levels of meaning along with tag type can be used to classify tags and assess quality; unfortunately, there is little agreement on a set classification of tag types. Sen *et al.* (2006) suggest three tag classes: Factual: *People, Places, Concepts*;

Subjective: *User Opinion* and Personal: *Organisation* and *Ownership*. Sen *et al.* (2007) and Al-Khalifa and Davis (2007) found that most users on del.icio.us enter Factual tags and few enter Subjective tags. Factual tags comprised 63% of their tag set, a finding mirrored by Sen *et al.* (2007) in their MovieLens dataset. Bischoff *et al.* (2008) define tag types as: *Topic, Time, Location, Type, Author/Owner, Opinions/Qualities, Usage Context* and *Self Reference*. They analysed a selection of tags from three different tagging systems to assess how tag type varied across tagging systems. Topic was the most used tag type in del.icio.us (web pages) and Flickr (images) and Type was the most used in Last.fm (music). Mathes (2004) provides eight tag categories: *Technical, Genre, Self-Organisation, Place Names, Years, Colours, Photographic Terms* and *Ego*. These categories are limited, predominately perceptual rather than conceptual, and unreflective of how users actually tag. Dubinko *et al.* (2007) suggest three tag categories: *Events, Personalities* and *Social Media*. Schmitz (2006) defines five tag categories: *Location, Activity, Event, People, Things*. Tag function is determined by how it is used not what it describes (Golder and Huberman, 2006; Strohmaier *et al.*, 2012). Golder and Huberman (2006) question whether the tag is descriptive of the resource itself or descriptive of the category into which it falls and suggest six tag functions: *What or Who the resource is about; What it is, Who owns it, Refining Categories, Identifying Qualities or Characteristics, Self-Reference* and *Task Organising*. Lin *et al.* (2006) summarises the tag categories presented by these researchers as *Location, People, Subjects* and *Event*. They define their own set of eighteen tag categories: *Place-name, Compound, Thing, Person, Event, Unknown, Photographic, Time, Adjective, Verb, Place-general, Rating, Language, Living Thing, Humour, Poetic, Number* and *Emotion*. The categories were evaluated using a Flickr dataset. They found that *Place-name* was used most (28%), followed by *Compound* (14%) and *Thing* (11%). Least used (<1%) were *Humour, Poetic, Number* and *Emotion*. This suggests a lack of motivation for Flickr users to use subjective or abstract tags and a predominance for basic level

descriptions, a finding also reported by Golder and Huberman (2006) on a sample of del.icio.us tags. They argue that tag sets converge with users quickly agreeing on basic level tags to describe similar resources.

System design affects user motivation to tag which affects tag type (Marlow et al., 2006; Bischoff et al., 2008). Conformity Theory (Sen et al., 2007) defines how social influence within a system affects how users tag, especially if system tags are made visible to the user - users will assign what they see (Golder and Huberman, 2006; Marlow et al., 2006; Sen et al., 2006). Cattuto *et al.* (2007) predict the rate at which users and the community re-use tags. Leysen *et al.* (2012) tested four conditions of suggested tagging: *no tags*, *correct generic tags*, *correct specific tags* and *incorrect specific tags*. Users entered more tags if no tags were suggested and these were significantly more generic. Specific tag visibility had a negative effect on the quantity of general tags produced. A similar finding was recorded by Arends *et al.* (2012): allowing users to view tags encouraged users to generate fewer tags of higher agreement. Stuart (2012) found that tagging images on Flickr tag type is dictated by content and not by social influence or user motivation. Arends *et al.* (2012) observed that more descriptive tags were assigned to popular images.

2.6.2 Motivation to Tag

Users tag either to organise their own personal collections, as a form of social interaction to improve visibility of their own resources or to disseminate resources they think will be of interest to the community, be those perceived or actual, for financial reward (e.g. Mechanical Turk) or for *Play and Competition* (GWAP) (Marlow et al., 2006). Personal perception of content also has a strong influence on motivation

to tag (van Velsen and Melenhorst, 2009; Arends et al., 2012). Freiburg *et al.* (2011) suggest that creation of new information is a motivation to tag. Personal and social motivation are categorised by Ames and Naaman (2007) and extended by Stuart (2012) as *Self-Organisation*, *Self-Communication*, *Social Organisation* and *Social Communication*. Stuart's (2012) investigation into user motivation to tag on Flickr revealed only two main motivations: *Self-Organisation* and *Social Communication*. In contrast, van Velsen and Melenhorst (2009) found that users in a custom video tagging system had low motivation for social communication and high motivation for social organisation.

Users incentivised by Self-Organisation are motivated to tag either by the resource or the system, for *Future Retrieval* (Marlow et al., 2006). They get most benefit from a system where they provide the resource (van Velsen and Melenhorst, 2009). A user's first compulsion to tag is self-focussed, they need to gain from it (*Personal Tendency* - (Sen et al., 2006)). Tagging becomes social once a user has spent time in the system and realised its usefulness (van Velsen and Melenhorst, 2009; Marlow et al., 2006), (*Community Influence* - (Sen et al., 2006)). Tags entered by users motivated by Social Communication are entered to appeal to a perceived or specific audience (*Contribution and Sharing* - (Marlow et al., 2006; Zollers, 2007; Melenhorst and van Velsen, 2010)). Social tags are only used in tagging systems where all tags are visible to all users. Stuart (2012) describe a 'selfishness' with Social Communication, where tags are entered to improve the visibility of one's own resources, to *Attract Attention* or as *Self-Presentation* (Marlow et al., 2006). Performance and Activism tags (Zollers, 2007) will be incentivised by this motivation. These are not useful tags, they can be playful but also malicious and they appeal to other users within a highly specific interest area. Social Communication can motivate users to enter *Opinion Expression*

tags (Marlow et al., 2006; Zollers, 2007) that can act as a review or recommendation of the resource (Zollers, 2007; Wang et al., 2012; Goh et al., 2011).

Strohmaier *et al.* (2012) suggest two types of tagger *Categorisers* and *Describers*. Categorisers are motivated by creating personal navigational collections of tags and resources (Self-Organisation - (Stuart, 2012; Ames and Naaman, 2007)). Describers are motivated by accurately describing the resource (Social Organisation - (Stuart, 2012; Ames and Naaman, 2007)). Describers are more likely to have unique tags and categorisers more likely to use tags with high agreement. Describers are more likely to be motivated to use a GWAP (Goh *et al.*, 2011); it would be pertinent to expect a higher level of unique tags in a game based tagging system and more convergence in non-game tagging systems. Arends *et al.* (2012) suggest that most users represent a mixture of Categorisers and Describers and tags demonstrate both motivations.

2-5 Summary of concepts describing motivations to tag.

Self Organisation	Future Retrieval (Marlow et al., 2006) Personal Tendency (Sen et al., 2006) Community Influence (Sen et al., 2006) Categoriser (Strohmaier et al, 2012)
Social Communication	Contribution and Sharing (Marlow et al., 2006) Attract Attention (Marlow et al., 2006) Self Presentation (Marlow et al., 2006) Performance (Zollers, 2007) Activism (Zollers, 2007) Opinion Expression (Marlow et al., 2006) Descriptor (Strohmaier et al, 2012)

2.6.3 Tagging Video

Ding *et al.* (2009) introduce the term 'Tagometrics' which is the process of understanding user tagging behaviour and usefulness of tags by counting tags and measuring frequency, co-occurrence and similarity. Many researchers believe that high frequency is the main measure of quality (Lin and Aroyo, 2012; Golder and Huberman, 2006; Cattuto *et al.*, 2007; Gligorov, 2012). Research into tagging video is sparse and the majority of it focuses on a quantitative approach using tagometric analysis rather than applying interpretative techniques used to analyse image tags.

Wang *et al.* (2012) suggest using both automatic and manual methods of tagging video so humans only tag what the computer cannot do. They define this as *Assistive Tagging*. As previously discussed, automatic annotation is not yet sophisticated enough to annotate with high level semantics (defined as '*Context Level*'), but is competent at assigning annotations of low-level features (defined as '*Content Level*'), albeit at high computational cost. Humans can tag both low and high level features but at a high labour cost. Combining the two methods could be a solution with general level tags being added automatically and humans adding the specific and abstract vocabulary. The problems for such a system are motivating users to tag and the computational cost of automatic methods. These types of tag recommendation systems are trained using human opinion on the relevance of the tags to the video. Gligorov *et al.* (2013) found that human raters rarely agreed on the relevance of a tag. Freiburg *et al.* (2011) and Ulges *et al.* (2008b) use concept detectors to aid users in tagging videos, suggesting tags to users that describe shots of their video will improve the training data for concept detectors and textual data and improve the precision of text-based search. This is in contrast to other researchers who use concept detectors to categorise videos for content search. Ulges *et al.* (2008b) note

that current tags on YouTube are weak, too few and do not describe individual shots of the video. Training on real world data could allow the system to be used on large corpora.

Velsen and Melenhorst (2009) suggest seven motivations to tag videos, derived from two focus groups. These can be split into Personal Indexing, Socialising and Communicating:

Personal Indexing:

- Re-find a video;
- Enable others to find a video;
- Clarify or add information;

Socialising:

- Find information related to the video;
- Recommend video to others;

Communicating:

- Find friends or like-minded people;
- Communicate with others.

A recent video tagging project, Waisda? has produced a number of studies on tagging video, focussing on how useful the tags generated are at improving video search. Several research papers have been published from the project: Oomen *et al.* (2010), Gligorov *et al.* (2010), Gligorov *et al.* (2011), Gligorov (2012), Lin and Aroyo (2012), Gligorov *et al.* (2013) and Hildebrand *et al.* (2013). Oomen *et al.* (2010) claim that social tagging stimulates active engagement with the content, it takes more time

to add tags to a video because of its temporal nature; gameplay can keep people engaged in the activity. In comparing tags to a professional cataloguing database and a Dutch version of WordNet, the authors found that only 2.7% of tags were present in both databases. Tags describe what a person sees, hears and feels, and relate to specific points in the video. Gligorov (2012) notes that tags are useful at describing objects in videos but found little evidence that they describe entire scenes. Similarly, professional annotations do not describe objects or scenes but categorise the whole video into topic based subject groups. As user tags consist of more descriptive vocabulary than categorising, the author poses the question of whether tags can be used to describe scenes.

Gligorov (2012) claims that tag quality is determined by how the tag will be used. In the Waisda? project, usefulness was defined by Oomen *et al.* (2010) as a tag that was similar to a professional annotation, making tags useful for curation but not description of content. They report that tag frequency does not correlate with usefulness, with only 2 of the top 20 most frequently added tags overlapping with a professional thesaurus. Lin and Aroyo (2012) report an increase in tag quality, with quality being defined as a tag that matches another player's tag between pilots one and two of the Waisda? project. Pilot two generated 51% matching tags whereas pilot one only generated 37%. Lin and Aroyo (2012) attribute this to improvements made to the interface in pilot two based on usability studies conducted during pilot one. Gligorov *et al.* (2013) elaborate on this finding, defining co-occurrence of tags as 'verified tags'. They hypothesised that tags upon which users agree will be more useful for search than unique tags. However, their research contradicted this. They found that verified tags had more general vocabulary and gave higher precision in search but lower recall, whereas tags only entered once had more specific vocabulary

and a higher measure of recall. Relying on tags with high agreement will not improve metadata as much as if unique tags are included as well. Hildebrand *et al.* (2013) note that user tags complement professional annotations by providing a user perspective of the content. This is important for online video collections as it is users that will search for the videos. Having metadata that uses terms similar to users' queries will improve search (Tjondronegoro *et al.*, 2009). Melenhorst *et al.* (2008) found empirical evidence that users employed the same terminology in the tagging process and the retrieval process and that user tags were more effective at video retrieval than professional annotations were.

There is an abundance of research on labelling images which is based on research on image interpretation. There is little similar research into interpreting video, in particular the cognitive difference between the two. Classification of video tags follows the same frameworks used to classify image tags. Whilst an image is static and therefore all tags refer to that one static image, video is temporal, consisting of multiple still images. Audio can be recorded as a tag as well as visual features. The cognitive cost of tagging a video has to be greater than tagging an image, but does this affect the types of tag users enter? Gligorov *et al.* (2010) compared tags to subtitles for a selection of videos. On average, one quarter of all tags could be found in the audio, rising to one third for verified tags. The authors attribute this finding to gameplay; users are incentivised by the scoring mechanism of matching with another user.

2.7 Summary

Section 1 highlighted the abundance of online video and distinguished UGC and PGC. A gap in knowledge was highlighted in how advertising affects user enjoyment of YouTube and whether users have a preference for UGC or PGC. A two-fold problem was acknowledged with video search. Current methods of video retrieval rely on text-based search which is insufficient at describing the video content yet current advances in content-based retrieval and automatic annotation only extract low-level features such as colour, texture and shape at a high computational cost. A possible solution and research area was identified to encourage users to annotate videos to improve textual descriptions of content. However, manual annotation involves high labour costs and is therefore not a cost effective solution. The proposed solution to this problem is to improve textual data using social tagging. However, no social tagging systems exist for video; YouTube only supports owner tagging. Agreement was found within the literature that YouTube tags were insufficient, but no detailed study on the types of tag in YouTube had been conducted. There are no attempts in the literature to investigate collaborative tagging for video and whether this is effective in describing the video content. Video tagging games were identified as the solution to labour costs and the lack of an acceptable video tagging system.

A question was raised by the literature review of whether a GWAP is a Gamified system. The answer has not yet been resolved by current work emphasising a requirement for further investigation. Through the discussion of GWAP design, it was apparent that little attention is placed on user enjoyment or user centred design. The majority of projects adapt the ESP Game model despite numerous flaws being highlighted. Little research has been conducted to investigate alternative GWAP design methods. Research into gamification begins to address this problem, but does

not go far enough to suggest game elements that are optimal for tagging games generally, or specifically video. Research into game design methods is extensive; the literature review revealed some key areas that could be adaptable to the design of a video tagging game, namely player types and fun factors. There is limited understanding of what motivates a user to engage in a video tagging game and how to design systems that will attract motivated users. The discussion of play and motivation theory provided insights so that understanding in this area can be developed further. The other notable omission in GWAP research is evaluation of tag output. This is mostly limited to measuring tag frequency as an indicator of quality, where high user agreement is equal to high relevance. Evidence in more recent research indicates that it is not always the case that high frequency tags are optimal for effectively describing the video content, although they may be optimal for search. A requirement was identified for a classification scheme to measure the descriptive quality of tags.

This research aims to fill the gaps in knowledge identified in this literature review, namely to:

- Investigate user motivation to participate in video tagging games;
- Provide design methods for the creation of video tagging games;
- Provide further insight into how users tag videos both in existing tagging systems and in video tagging games;
- Discover, through the application of a tag classification scheme, whether game elements can be used to encourage users to tag and enter tags of differing descriptive quality.

elements chosen need to motivate certain player types and incentivise users to move past the 'just try it out' motivation and to sustain use. Methods to measure user experience and user engagement will be employed to evaluate the success of the design, focussing on which game elements users enjoy.

R2. Can game elements affect the types of tag that players enter?

Previous GWAP research has not analysed the descriptive quality of the tags produced, with most equating tag quality with high agreement. Two exceptions are Gligorov *et al.* (2013) and Goh *et al.* (2011), who both evaluated tag quality based on basic level theory. This research uses a custom classification scheme adapted from the literature to evaluate the language of tags, their relevance and how they describe the video content. There is no research that compares game based tagging with non-game tagging of videos. A small scale study comparing non-game and game based image tagging conducted by Goh *et al.* (2011) found that game based tagging increased tag quantity, but that tag quality was higher using the non-game tagging system, with more matching (basic level tags) generated in the game environment. Related to this, the research will explore users' preference for a game environment over a non-game environment. If the game elements are stripped leaving only a simple video tagging system, will users be differently motivated to use the system and will it affect the types of tag they enter? The research will compare the tags generated using VideoTag to those assigned to videos on video sharing sites, investigating whether the extrinsic motivation provided by game elements, compared to the intrinsic motivations to share and organise, affects the types of tags entered. The analysis will consider whether users need the game elements to generate useful data and also how individual aspects of gameplay can affect tag type.

R3. Does video content affect tag type?

Previous video tagging GWAP projects have focussed on labelling professionally generated videos. This research concentrates on encouraging tags for both user generated and professionally generated YouTube videos. Previous research (Greenaway, 2007) assumed that including only comedy videos in the tagging games would make them more entertaining. This assumed all users would prefer to tag comedy videos and did not account for the fact that preference for certain content has an impact on enjoyment. Hildebrand *et al.* (2013) found that the popularity of a video had an impact on the quantity of tags, but tag quality was not assessed. Stuart (2012) found that the content of images affected tag type more than a user's intention to tag. Experiments will explore whether the general category of video has an impact on user motivations to tag and whether differences in video content change the quantity and the types of tags users enter. Further investigation will question whether video content aimed at special interest groups will have an effect on the specificity of language that users use when assigning tags.

R4. Can video tagging games encourage users to enter specific level descriptive tags as well as general level descriptive tags?

The literature review highlighted how using the ESP Game strategy of scoring upon agreement will produce more basic level tags, revealing a requirement for users to be encouraged to enter tags at varying levels of abstraction. In an owner tagging system tags rely on the expertise of one user, but in a collaborative tagging system the expertise of many users appears in the vocabulary, potentially giving a broader range of tags. YouTube videos are indexed not by their content but by the textual

data assigned to them or surrounding them when embedded in a webpage. Improved textual data should include collections of keywords or multiple word phrases that describe the video content accurately, capturing objects and actions within the video and interpreting content and expressing opinions. This research experiments by adjusting the gameplay to determine whether more specific level and subjective tags can be encouraged. The absence of any personal or social motivations to organise or communicate using tags should encourage users to only enter tags that describe the video. When using tags to improve textual data a tag in isolation cannot be measured for quality or usefulness until it is compared to other tags assigned to the same video. The quality of a set of tags is the extent to which it contains a range of general and specific objective language that accurately describes the content of the video. Individual tag quality relates to how it contributes to the tag set. In this research tag usefulness is not measured. A tag is assumed to be useful if it is relevant to the video content and does not refer to system properties or have a social or organisation function. The tag analysis will centre on whether video tagging games are a reliable method for generating improved textual data for online videos.

4 Classification Studies – Straight Tagging of User Generated Video

4.1 Tagging YouTube

4.1.1 Introduction

The literature review revealed a gap in knowledge in how users tag on YouTube and the types of tag they use. This section presents an analysis of tagging behaviour on YouTube, through a classification of the user-generated tags assigned to a random selection of 100 YouTube videos. Tags were classified into various categories of tag type, using a custom classification scheme. This is designed to help understand which attributes comprise a useful tag. The work is a preliminary study to gain a more detailed understanding of the tagging behaviour of YouTube users and the types of tag they enter, extending the findings of Ding *et al.* (2009). Investigation into the tag vocabulary that exists on YouTube will highlight any improvements that could be made.

4.1.2 Literature Review

YouTube provides User Generated Content (UGC) to mass audiences. The diversity of user generated video creates difficulties for categorisation and findability (Yang *et al.* 2007). At present, tagging on YouTube is not collaborative, with only the owner of the video being able to tag. If collaborative tagging was introduced, any user could tag any video; there would then be the potential for more tags to be entered and for rich folksonomies to be created. Collaborative tagging can be useful for improving indexing and categorising internet video by providing a user generated text alternative to the visual content. Collaborative tagging as a low cost method of video annotation has value because the viewers of the video create it. The tags may describe the content or express opinions about the video. Collaborative tagging can

also provide a social commentary about a video (Shamma et al. 2007). With the potential for international audiences, multilingual tags can also make a video more widely accessible. Tags can be unbiased; therefore, unlike keywords generated by the creator of the video, they offer a reflection of the content of the video from a wider range of perspectives.

Halvey and Keane (2007), in a study of YouTube, found that more descriptive information about a video correlated with more views. This is explained by the fact that search engines use text matching techniques to find videos; therefore, the more textual information surrounding the video, the higher the probability of it being returned by a query. Alternatively however, better videos may just be more extensively tagged. This supports the idea that increasing the number of tags for a video will improve video search. The cheapest way to increase the number of tags is to introduce collaborative tagging. Geisler and Burns (2007) published findings of a quantitative analysis of YouTube tags. They found the mean number of tags per video to be 6 and that 66% of the tags added additional description of the video content that was not found in other text on the page, such as the title, description or author. Halvey and Keane (2007) conducted a study of YouTube focussing on search and user behaviour. They found that most users only use YouTube to search and watch videos, but few users interact with the social element of the site i.e. join groups, upload videos, make friends, favourite videos or comment. If collaborative tagging was implemented on YouTube, passive users would benefit from the tags entered by the active users. A further study found that videos with few tags received few views. The average number of tags per video, for their dataset, was 4.1 with the maximum amount entered being 25. However, for videos that are recommended by YouTube on the front page, the average number of tags was found to be double that

at 8.73%. The authors found no evidence that considerably increasing the amount of tags beyond the recommended video average substantially increased views. Ding et al. (2009) analysed tags and tagging behaviour as part of a comparison study of social tagging over three social networks, del.icio.us, Flickr and YouTube. The popularity of tags over time was analysed by comparing the most popular tags and tagging behaviour over two years, 2005 and 2007. By comparing tag popularity over time emerging trends in topics of interest were revealed. The study highlights a problem with analysing tags in YouTube, as because only the user uploading the video can tag, there is no indication of the collaborative opinion of viewers of the video. YouTube tags can only indicate trends in the type of content being uploaded to the site, but cannot offer insight into the type of content users prefer watching. The authors note that using tag frequency to identify community interest is not possible in YouTube.

4.1.3 Research questions

A study was conducted to facilitate further research into video tagging games as a suitable method of collaborative tagging. The aim was to improve understanding of what attributes are useful for tags in terms of improving textual descriptions of user-generated video. The following two research questions are addressed:

1. How useful are the tags entered by the uploader of a video at describing the content to other YouTube users?
2. Does the absence of collaborative tagging impact on the common types of YouTube tag used and the cognitive level of the tag vocabulary?

4.1.4 Methods

4.1.4.1 Data Collection

The dataset of Ding *et al.* (2009) was used for this study. The data was originally collected as follows: In September 2007 a crawl of YouTube was conducted to obtain a dataset of video URLs and tagging data. The crawler started from the main page at <http://youtube.com> and visited every available video page (links starting with <http://www.youtube.com/watch?v>). On each video page it collected tagging data and visited the links pointing to other video pages. YouTube does not provide related tag data. In order to avoid visiting the same page more than once, the query parts of links were ignored (i.e. <http://www.youtube.com/watch?v=X2IExa2A198> and http://www.youtube.com/watch?v=X2IExa2A198&watch_response lead to the same video).

The original dataset contained 43,641 tags. The majority of non-English words or characters (Chinese/Japanese) that had not converted correctly into the text file were manually removed; 1,461 entries were removed leaving a dataset of 42,180 tags. A random selection of 100 videos and their assigned tags were then extracted from the dataset using a custom script. This created a dataset of 768 tags for classification.

4.1.4.2 Classification Scheme

Angus *et al.* (2008) developed a classification scheme based on possible image categories in Flickr, notions of “of” and “about” (Shatford, 1986, 1984 in Angus *et al.*, 2008) and the notions of tag type defined in Golder and Huberman (2006). Categories in the scheme were further grouped based on social or personal motivation to tag.

For the purposes of this research, the classification scheme was modified to be more suited to a classification of YouTube Tags. As tagging on YouTube is primarily socially motivated and carried out by the uploader of the video, there was no requirement for the task organising category (e.g., tags such as toread, toprint, towatch). Angus *et al.* (2008) found no task organising tags in the Flickr study and an assumption was made this would also be true of YouTube tags. The distinction between social and personal motivation was removed, with categories in A and B being tags generally descriptive of the content and categories in C being of use only to specific users or groups within the YouTube community. Rather than miscellaneous categories as defined by Angus *et al.* (2008), categories in D are tags which are either irrelevant, or seen as not useful in terms of describing or indentifying the video in search or tag browsing.

Alongside restructuring the classification scheme, five new categories were added. With the addition of category A2, a distinction was made between tags that identified generically what the video is of and that identified a YouTube category or asserted a genre (e.g., Comedy, Music, Horror, Rock). Category B2 was created for tags that expressed an opinion about the video as a whole or certain qualities and characteristics of the video, such as funny and scary. Three categories were added to account for irrelevant tags. D2 (multi-word tags) handles tags that only have meaning when viewed in context with the other tags assigned to the video. Often names, titles and descriptions are entered as tags, but as single word tags viewed in isolation the tag becomes meaningless. This practice leads to an abundance of conjunctions and prepositions (e.g., the, in, of, and) and a separate category D7 was created to handle these tags. Category D3, attention attracting tags, was added from an assumption that some users uploading videos have a primary motivation to tag to

get more views for their video and would therefore add tags containing popular search terms (e.g., porn, sex, celebrity name) in order to achieve this. Only one classifier was used therefore the findings are limited to the view of this classifier. Table 4-1 below shows the classification scheme used and explains each category.

Table 4-1 The tag classification scheme including category definitions - adapted from (Angus et al. 2008).

A		Generic relationship between tag and video content
	1	Tag identifies what video is of at its most primary and objective level - no subject specific knowledge is needed to make this distinction (e.g., a video of a cat, tagged as 'cat' or 'animal'). Also included is the tag video.
	2	General YouTube defined Category or Genre (e.g., Comedy, Entertainment, Music)
B		Specific relationship between tag and video content
	1(a)	<p>Tag identifies what video is of. Familiarity or some existing knowledge is needed to make this connection, and to a certain extent an assumption has to be made about this connection.</p> <p>Tags which identify place names/events – a video of a concert tagged with the band name and venue, or a football match tagged with the team name, or an individual's holiday video tagged with the destination, requires knowledge acquired from familiarity with the specific place/event in question. Assumptions have to be made that a video tag is what it claims to be if the video is not familiar.</p> <p>Tags which identify people/animals/objects – a video of Elvis Presley tagged as 'Elvis Presley' requires knowledge and familiarity of Elvis</p>

	1(b)	Presley. Distinctions cannot always be made between 'famous' people and 'non-famous' people, therefore the assumption has to be made that a video of a girl, tagged as 'Sarah' is a video of a girl who is called 'Sarah'.
	2	Tag identifies what the video is about Typically expressed by the use of abstract nouns or adjectives - an interpretation is made of what the video is about (e.g., video of people smiling tagged as 'happiness'; video of cars on a motorway tagged as 'speed').
	3	Tags which express opinion of the content Includes Golder and Huberman (2005) tag types of <i>Qualities and Characteristics</i> and <i>Opinion Expression</i> (e.g., 'funny', 'rubbish')
C		Tag only useful to a minority of users, specific individual or group
	1	Refining tag Tag which cannot stand alone - only useful when looked at as part of the larger tag set (e.g., episodes of a series of videos specified by a number; acronyms or dates.).
	2	Self-reference tag Tags which identify video content in terms of its relation to either the tagger or the specific group which the video belongs to (e.g., 'my dog'; 'our graduation') OR tags which appear useful, but show no relationship/connection to the accompanying video.
	3	Tag which explicitly denotes ownership of video (e.g., video tagged with the same username as that of the person who uploaded the video).
D		Irrelevant/Non Useful Tags
	1	Compound tag - Tags where words, phrases and sentences are joined together as one long text string.
	2	Multi-word Tags - Tags that as single words are meaningless, but placed in context with the other tags have meaning. (e.g., Celebrity name, Title

		of film, TV show, song, video game)
	3	Attention Attracting Tags – Tags that are assigned to attract attention to the video, that refer to popular search terms, but have no relevance to the video content. (e.g., Porn, Sex, Celebrity name.)
	4	Misspelling (e.g., ‘Belguim’ instead of ‘Belgium’) Whilst it may be obvious what the tag is meant to be, a misspelling obviously renders the tag useless in terms of subsequent users of the system who are searching for videos with that specific tag, unless they too misspell the tag/word.
	5	Unable to determine relationship Despite having attempted to look up either the meaning of the tag and whether the tag is a foreign word or not, tags which do not fit into any of the above categories will be deemed as unable to classify (e.g., nonsensical words).
	6	Foreign word/character
	7	Conjunctions and prepositions (e.g., the, in ,of, and)

4.1.4.3 Findings

It was important to classify the tags whilst watching the associated video in order to correctly ascertain the meaning of some of the tags. For instance, it is difficult to classify C3 (denotes ownership) tags without first visiting the video page to find the username of the uploader. This fact questions the usefulness of some tags, as to be useful for search and discovery of video, they need to be meaningful in isolation from the content. Some videos were no longer available, and so the tags assigned to these videos were classified into the D5 (unable to determine relationship) category.

Despite it being possible to classify some of the tags, a decision was made that the tags could not be accurately classified without watching the video first.

A large number of tags referred to people: some were famous people and some were people in the video, the creator, or the username of the uploader. This is not reflected by the B1b (people/animals/objects) result of 9.5% as the majority of these tags were classified into the D2 (Multi-words) category. The largest percentage of tags, 23.3%, were placed into the D2 category. Some of the tags classified in this category resulted from complete sentences being placed in the tag field, either as a description of the content or the title. The majority, however, were names of people, bands or album titles that had been entered as two or more words. Considering this tagging practice by users, a surprisingly low result of 3.3% was recorded for the D7 (Conjunctions and Prepositions) category. It was expected that a higher percentage of these tags would be found in relation to the other categories, due to the finding in Ding *et al.* (2009) that 'the' is the most frequently assigned tag for the years 2006, 2007, and fourth in 2005. Analysis of the dataset of 100 videos used for this research revealed that 'the' constituted 1.4% of all tags and was also the most frequently used tag in the dataset.

Table 4-2 Total number of tags and corresponding percentage of all tags.

Classification Category		No of tags	Percentage of all tags
A1	Tag generically identifies what video is 'of'	85	11.1%
A2	Tag identifies video Category/Genre	42	5.5%
B1a	Tag specifically identifies what video is 'of'	66	8.6%

	(place names/events)		
B1b	Tag specifically identifies what video is 'of' (people/animals/objects)	79	9.5%
B2	Tag identifies what video is 'about'	67	8.7%
B3	Tag identifies opinion expression	51	6.6%
C1	Refining tag	45	5.9%
C2	Self-reference tag	5	0.7%
C3	Tag which explicitly denotes ownership of video	8	1%
D1	Compound tag	3	0.4%
D2	Multi-word tags (individual words in these)	179	23.3%
D3	Attention attracting tags	3	0.4%
D4	Misspelling	4	0.5%
D5	Unable to determine relationship	39	5.1%
D6	Foreign word/character	67	8.7%
D7	Conjunctions and prepositions	25	3.3%

These findings suggest poor tagging practices for many YouTube taggers and highlights that there is no shared vocabulary for tagging or a tagging standard as found in other systems like del.icio.us or Flickr (Ding et al., 2009). This is further highlighted by the lack of compound tags found in the dataset; only 0.4%. In contrast, Angus et al. (2008) found 12% of the tags in the Flickr data sample to be compound tags. However, a possible reason for the large percentage of compound tags is that Flickr handles multi-word tags by converting them to a compound tag. YouTube has no system in place to try and encourage useful tags either via suggestions as in del.icio.us, or converting the user inputted text into a more usable style, like in Flickr. These findings reflect the continued vocabulary problem faced by all tagging systems (Furnas et al., 1987; Golder & Huberman, 2005).

Category A1 (what the video is of) and A2 (category/genre) contain mostly basic level tags that describe the content at its most general. 11.1% of all tags were classified A1, the second highest category. Surprisingly, A2 contained only 5.5% of tags, suggesting that YouTube taggers describe the video content more than they use tagging to categorise the video, using the pre-assigned YouTube categories only. This finding is emphasised by the high percentage of Category B tags, that more specifically describe the video content and may require some specialist knowledge. B1b (9.5%), B2 (what the video is about) contained 8.7% of tags, B1a (places/events) contained 8.6% and B3 (opinion expression) 6.6% of all tags. Further indication that YouTube taggers use more specific level vocabulary over basic level generalised terms is that 5.9% of tags were classified as C1 (refining tag) tags. The tendency of YouTube taggers to use more subordinate level, descriptive tags could explain the low percentage, 0.4% of category D3, attention attracting tags. It would be expected that these tags would be of basic level vocabulary, maximising the probability of agreement on terms, with tags being words that are perceived to be regularly searched for, or relate to popular categories or videos. To accurately assess the specificity of the tag vocabulary, tag frequency and co-occurrence metrics can be analysed (Golder & Huberman, 2005; Cattutto, 2007). This is not useful with this data sample as only 6.6% of tags occur more than once.

Despite having manually removed the majority of non-standard English characters from the database, some foreign words using standard English characters were overlooked (8.7%), the joint fourth most common tag category. In retrospect, if all foreign words had been left in the dataset a more realistic gauge of non-standard English tags in the YouTube system could have been discovered. This would have

been useful to indicate the international appeal of the YouTube website and the variety of content. It would give weight to the concept that the YouTube system would benefit from collaborative tagging as multi-language tags can help make the videos cross language barriers and be available for viewing by a wider audience.

4.1.5 Discussion

Collaborative tagging allows the taggers in the system to classify and categorise the content in the system using language useful to the community. In YouTube this doesn't exist, as only the owner of the video can tag and they may not use language or a style of tagging that is useful to the community. Without collaborative tagging there is no agreement between taggers that tags are good, useful and relevant to the content. In a collaborative tagging environment taggers will reuse tags they think describe the content well, or are useful to their purpose and a standard is created for tag vocabulary in the system i.e. truncating or compounding names to form one tag, rather than two individual, not so meaningful tags (e.g., russell-crowe, russellcrowe, russell, crowe). YouTube does not support this. More multi-word tags were identified than compound tags. Multi-word tags may be meaningful when displayed with other associated tags, but not in isolation. This renders them less useful for search, or browsing through tags. Whilst compound tags can be significant if seen in a tag cloud and could be used to browse tags to find videos, they are not useful for search as users will enter either single or multiple keyword searches. This creates the problem of how to accept and handle multi-word tags in a tagging system.

The classification suggested that the majority of YouTube tags in this dataset were of a subordinate level. Whilst these tags may be useful at finding less popular videos

through keyword search, in theory, searchers are unlikely to use more specific vocabulary for keyword terms, so the tags may well be relevant to only a few users rather than the majority (Furnas et al., 1987; Golder & Huberman, 2005). It could be the case that the random sample did not collect many videos with similar content, explaining why there were such a high percentage of tags that only appeared once in the dataset. Assuming the result is reflective of the YouTube system as a whole, if collaborative tagging was introduced, the percentage of tags that occur once might be reduced as more users entered tags that described the content? It may not be the case that the syntax used is too specific for the majority of users, but rather that without the collective vocabulary provided by collaborative tagging it is impossible to accurately assess the specificity of the tags or the level of agreement of terms achievable. The lack of agreement between YouTube tags makes the clustering of videos for related content impossible, impacting on their potential for categorising user-generated videos.

4.1.6 Conclusion

The results suggest that YouTube users use tagging as an extension of the description and title fields. Tags do not appear to be used to further categorise a video, with users apparently relying on the categorisation structure of the YouTube system for this purpose. This is surprising since Flickr tags seem to be frequently useful for this purpose (e.g., Angus et al., 2008) and suggests that YouTube video posters are less aware of the need to publicise their work through tags. The classification found that YouTube taggers used a relatively specific vocabulary to describe their videos, for instance, tagging the species of dinosaur, rather than just tagging dinosaur; or tagging the make and model of motorbike, as opposed to just entering the motorbike tag. These tags are useful to a minority of users, as the majority of YouTube users

probably want to be entertained, rather than to use the system to find specific video contents.

Through analysis and classification of collaborative tagging data it is possible to evaluate the collective intelligence of the community, to assess the social impact of a resource or user, to discover community interest, trends, popularity and social connections. The method of tagging implemented in YouTube does not allow for such evaluations, and it is not clear why this is the case. With the introduction of a collaborative tagging system it would be possible to assess the popularity of the videos through analysis of the amount of tags entered per video, the type of tag entered, language used and opinions expressed. Trends in viewing habits could be uncovered, which could improve the recommendation of videos. Recommendation systems could be developed based on shared user interest and co-occurrence of tags. The tags themselves could provide a method for categorising the increasing amount of user-generated content, either for retrieval, for curating collections, or for preservation of content.

N.B. This section was published and presented as a research in progress paper at ISSI 2009: (See Appendix A)

Greenaway, S., Thelwall, M., & Ding, Y. (2009). Tagging YouTube - a classification of tagging practice on YouTube. Proc. 12th International Conference on Scientometrics and Informetrics, 14th-17th July, Rio De Janiro, Brazil. P.P.660-664.

Greenaway is the author's previous name.

4.2 Broad Video Tagging

4.2.1 Introduction

There are two main types of tagging system: owner tagging, which produces a narrow folksonomy, and collaborative tagging, which produces a broad folksonomy. The previous section discussed a classification of tags from YouTube, an owner tagging system. This section presents a classification of tags collected from both YouTube and Viddler⁵ which incorporates collaborative tagging. Since this study was conducted in 2010 the Viddler system has changed; it no longer supports UGC, is not free to use and has adopted more of a business-to-business model. The YouTube model has also changed with less emphasis on categories and more on individual user or company channels. The systems discussed in this section refer to their structure in 2010.

Collaborative tagging in Viddler allows the taggers to classify and categorise the content in the system using language useful to the community. Viddler rely on the tags users generate to categorise their video library, in contrast YouTube users who rely on their own pre-defined categories, with tags being used only as additional textual descriptions of the video. Users have little incentive to tag their videos and tagging is not actively encouraged. Viddler users are motivated to tag videos for personal and social organisation (Van Velsen and Melenhorst, 2009). Users of Viddler have incentives to enter more tags per video than users of YouTube. The different levels of perceived usefulness provide users with different incentives to tag; this could affect the amount and the quality of the tags. The types of tags users enter will

⁵ <http://www.viddler.com/>

be affected by whether the tags are intended to categorise or describe the video (Strohmaier et al., 2012). The previous preliminary study of YouTube tags found that YouTube taggers use relatively specific vocabulary to describe their videos (see Section 3.1.1). The structure of Viddler may encourage users to take the role of categoriser over describer.

Halvey and Keane (2007) found the number of tags per video increases if the number of views is higher. Users in collaborative tagging systems will tag videos that are already tagged or popular (Sen et al., 2006; Arends et al., 2012). This section investigates whether these trends are evident in the Viddler and YouTube dataset and if there is any evidence of collaborative tagging in the Viddler dataset. A further aim of the study is to discover if the language of the tags or tag type is different depending on whether the tags were entered into a broad (Viddler) or a narrow (YouTube) tagging system. In addition, the research analyses whether the tag type is affected by the category of the video.

The following research questions are addressed:

1. Is there evidence of collaborative tagging activity on Viddler?
2. How does tag type differ between YouTube and Viddler?
3. Does the category of video affect how the video is tagged both in amount of tags and descriptive quality?

4.2.2 Methods

YouTube organises its video library into separate pre-defined categories whereas Viddler organises its video library using tags only. Eight categories were selected from YouTube, four categories that could be classed as entertainment and four categories that could be classed as informative. For each YouTube category the respective tag was then found on Viddler (see Table 4-3). Twenty videos and their tags were extracted from each category, from both YouTube and Viddler. The YouTube category of 'Gaming' did not exist on Viddler so the tags 'Game' and 'Games' were used instead. In this instance, ten videos were extracted for each tag.

Table 4-3 Category Groupings.

Category Type	YouTube	Viddler
Entertainment	Comedy	Comedy
	Entertainment	Entertainment
	Gaming	Game/Games (10 from each)
	Music	Music
Informative	Technology	Technology
	Sport	Sports
	Travel	Travel
	News	News

4.2.3 Data Collection

The YouTube API and Viddler API were used to capture a dataset of unique videos and associated textual data (title, user, views and tags). A custom PHP script was created to parse RSS feeds from each site and record the data in a database. For YouTube, 300 of the 'most recent' videos in each category were retrieved each day over a five day period (8th-12th Feb 2010). No duplicates were collected so it was not possible to collect 300 videos from each category, each day. For Viddler the feed was

ordered by 'most recent', all videos were retrieved in each of the 8 categories on one day (8th Feb 2010). The dataset was cleaned, removing videos that had been deleted from the system, or contained 100% non-English word/character tags. This left a YouTube dataset of 10,870 unique videos and a Viddler dataset of 46,573 unique videos, (see Table 4-4).

Table 4-4 The breakdown of videos over the eight categories.

Category	Total no. of videos	
	YouTube	Viddler
Comedy	1564	6119
Entertainment	2049	2954
Gaming	2067	10534
Music	1636	11509
News	1321	4992
Sport	1696	2803
Technology	1633	2458
Travel	1514	5204

To select videos for the classification study, 100 videos from each of the 8 categories were selected at random. A resample was taken for videos that had been removed from the YouTube system. To select the tags 1 tag per video was randomly extracted from each category, for both datasets, giving a total of 800 tags for each dataset. The custom classification scheme used in Section 4.1 was used to classify the tags for tag type and language parameters. Two additional categories were added to the classification scheme based on observations during the YouTube study Section 4.1; D8 (repeated tag) and D9 (URL).

4.2.4 Findings

4.2.4.1 Comparison of Tagging Systems

The highest number of tags per video on YouTube is 82, on Viddler 1701. The mean number of tags per video for YouTube was $M=12.42$, $SD=12.99$, $N=10,874$ and Viddler was $M=17.81$, $SD=22.54$, $N=46,574$. This contrasts with Halvey and Keane (2007) who report means between 3.7 and 8.7 for YouTube and Geisler and Burns (2007) who report a mean of 6.

The majority of videos have 4-20 tags on Viddler; the highest frequency of tags per video is 8 (3,262) followed by 9, 10, 11, 12 (2261), there is a considerable drop in frequency over the small range (see Figure 4-1). In contrast, most videos on YouTube have 1 tag (794), followed by 3, 6, 2, 5 and 7, with frequency dropping to 663 (see Figure 4-2). Taking the twenty videos with most tags, the Viddler set contains 77.1% of all videos, YouTube only 19.7%. These findings emphasize the more even distribution in the YouTube dataset than in Viddler, with more consistency in the amount of tags per video. An anomaly in the YouTube dataset was that 125 videos have a tag frequency of 68; inspection of the raw data showed the videos were all uploaded by one group of users all using the same tags for a large selection of videos with a similar theme. An anomaly in the Viddler dataset was also found; there is only one video containing 2 tags, compared to 671 containing 1 tag and 856 containing 3 tags. There are a large number of videos with similar tag frequencies on Viddler; this could highlight a convergence of user tagging behaviour or a standard user's follow when tagging or it could be the result of an automated method of entering tags. Unfortunately the Viddler data accessible through the API does not reveal which user entered the tags. Without this knowledge, it is impossible to discover through user tagging behaviour whether collaborative tagging exists.

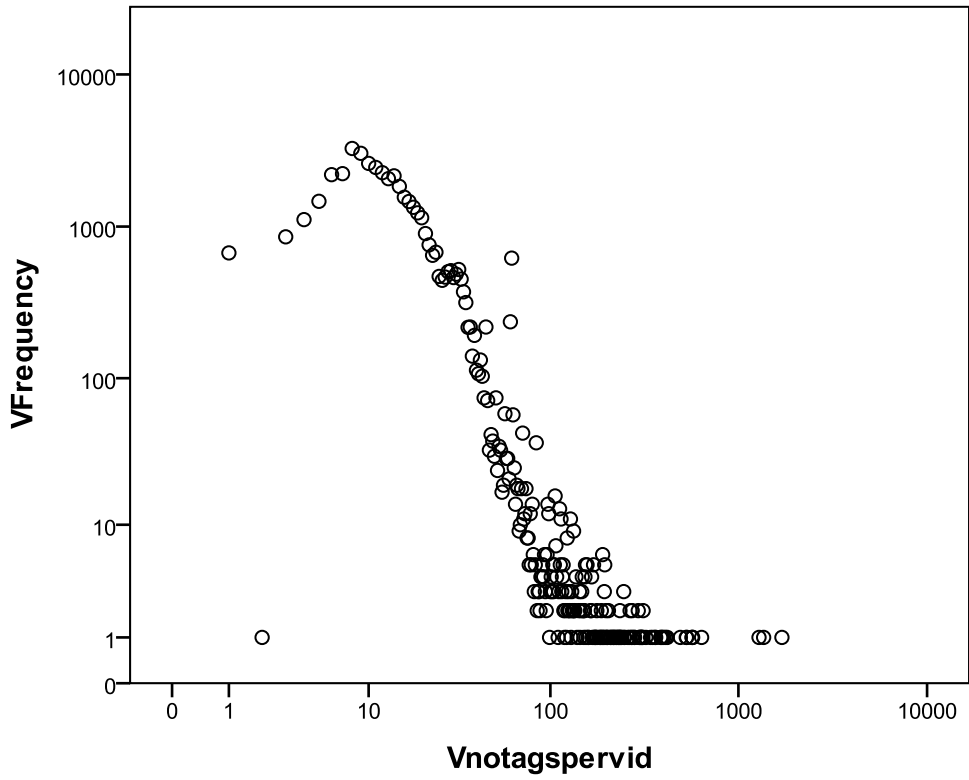


Figure 4-1 - Frequency distribution of tags per video on Viddler (log-log scale).

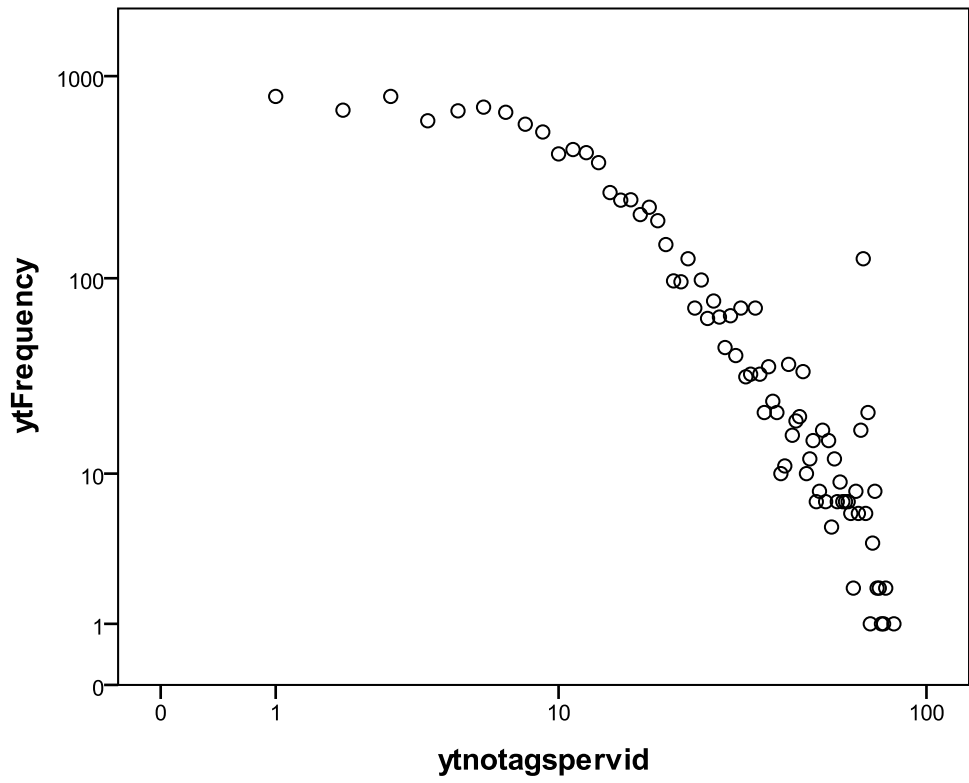


Figure 4-2 - Frequency distribution of tags per video on YouTube (log-log scale).

Halpin *et al.* (2007) suggest that a power law in tag frequency distribution indicates collaborative tagging. The tag frequency distributions in both datasets (Figure 4-1 and Figure 4-2) are not different enough to suggest different tagging practice between systems. Both very approximately conform to a power law, however even though collaborative tagging is not offered by YouTube. Although it is not possible to compare whether collaborative tagging has an effect on the amount of views a video receives, the relationship between views and tags can be analyzed. The mode number of views of videos on both YouTube (3.7%) and Viddler (8.6%) is 0; the majority of videos in both datasets have a low number of views, few have a high number of views. This could be attributed to the 'most recent' setting used for data collection. The means were distorted by a high number of views for a few videos in YouTube. The amounts of views per video on Viddler were considerably less than on YouTube (Figure 4-3). The data shows a few anomalies at both ends of the scale, with a few videos having low views but a high amount of tags, and videos with high views but a low amount of tags. The majority of videos had a similar amount of tags per video and views per video. Viddler had on average more tags per video than YouTube, but less views per video. The lack of views is explained by a more limited audience than YouTube can command. The amount of tags however cannot be empirically explained by the collaborative tagging system as it is not clear whether tags are entered by the uploader (owner) of the video, or by the owner and viewers.

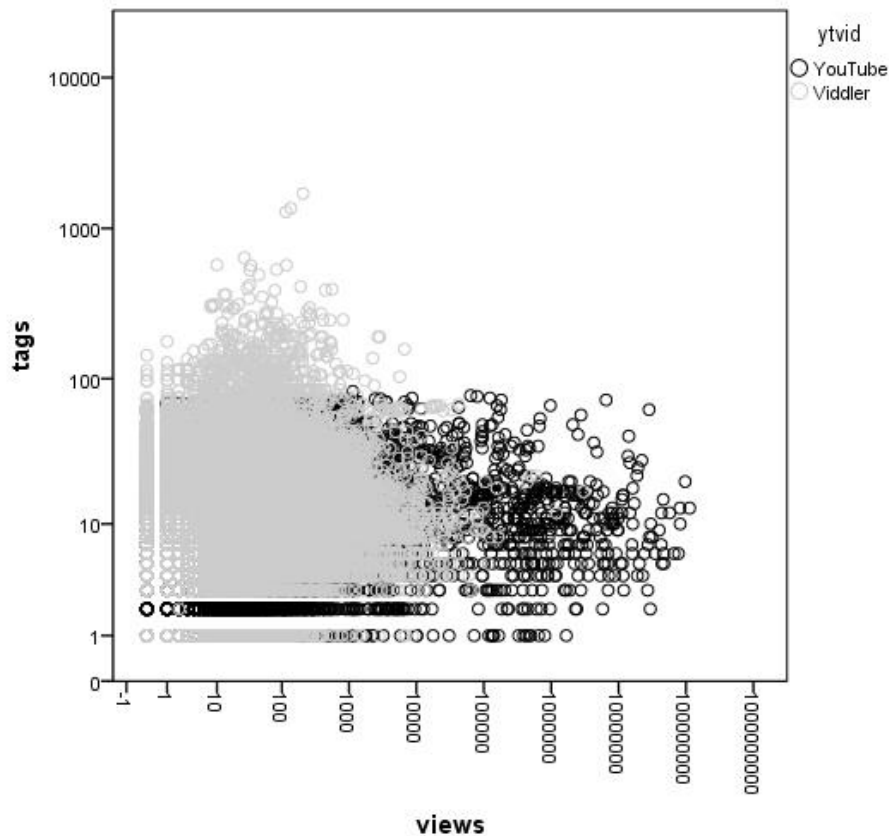


Figure 4-3 Comparison of tags and views on YouTube and Vidler.

H0 There is no relationship between the number of tags and number of views on YouTube and Vidler.

Weak relationships were found between views and tags for YouTube and Vidler using Spearman's correlation test. There was a weak positive correlation between views and tags in YouTube $R_s=.252$ and a very weak negative correlation in Vidler $R_s=-.035$, both are significant at $p=.000$. There is a small probability that assigning more tags to a video in YouTube will increase views. In Vidler, where tags are the sole method of categorization, assigning a large number of tags might decrease views. This shows that assigning more tags to a video does not guarantee higher

views; other metrics must also affect popularity. The results support Halvey and Keane (2007), who found evidence suggesting videos with higher views contain more tags, although their findings were not statistically tested and these findings suggest only a weak relationship. The null hypothesis can be rejected for YouTube and Viddler.

4.2.4.2 Tag Classification.

Since there is no clear evidence of collaborative tagging on Viddler, it is not possible to compare tag type between collaborative and owner tagging systems. Instead, tag type can only be compared between Viddler and YouTube. The tag classification focuses on differences in tag type between different categories of video in each system and between the entertainment and informative category groupings. Figure 4-4 shows the distribution of tags over the eight categories. The YouTube Technology, Gaming and Entertainment categories have the largest proportion of tags; in Viddler Gaming and Music have the most tags, with few tags in Sport or Technology.

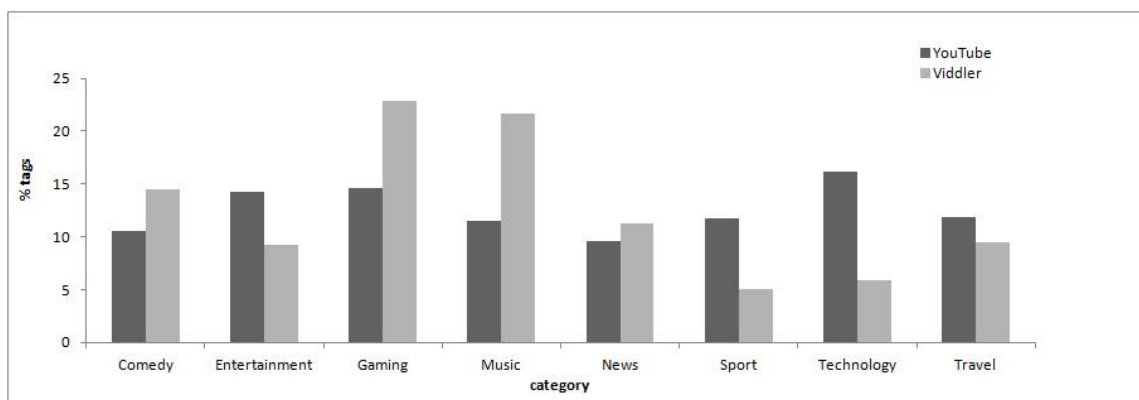


Figure 4-4 A comparison between YouTube and Viddler tags.

Table 4-5 shows how the tags are dispersed over the two category groupings; Viddler has considerably more entertainment tags than informative tags whereas YouTube tags have a fairly even split between the two categories. Viddler has more entertainment tags but less informative tags than YouTube.

Table 4-5 The percentage of tags assigned to Entertainment and Informative videos in YouTube and Viddler.

Video Category	YouTube	Viddler
Entertainment	50.8%	68.3%
Informative	49.2%	31.7%

Results of the tag classification are presented in two graphs separated into entertainment (Figure 4-5) and informative (Figure 4-6) categories.

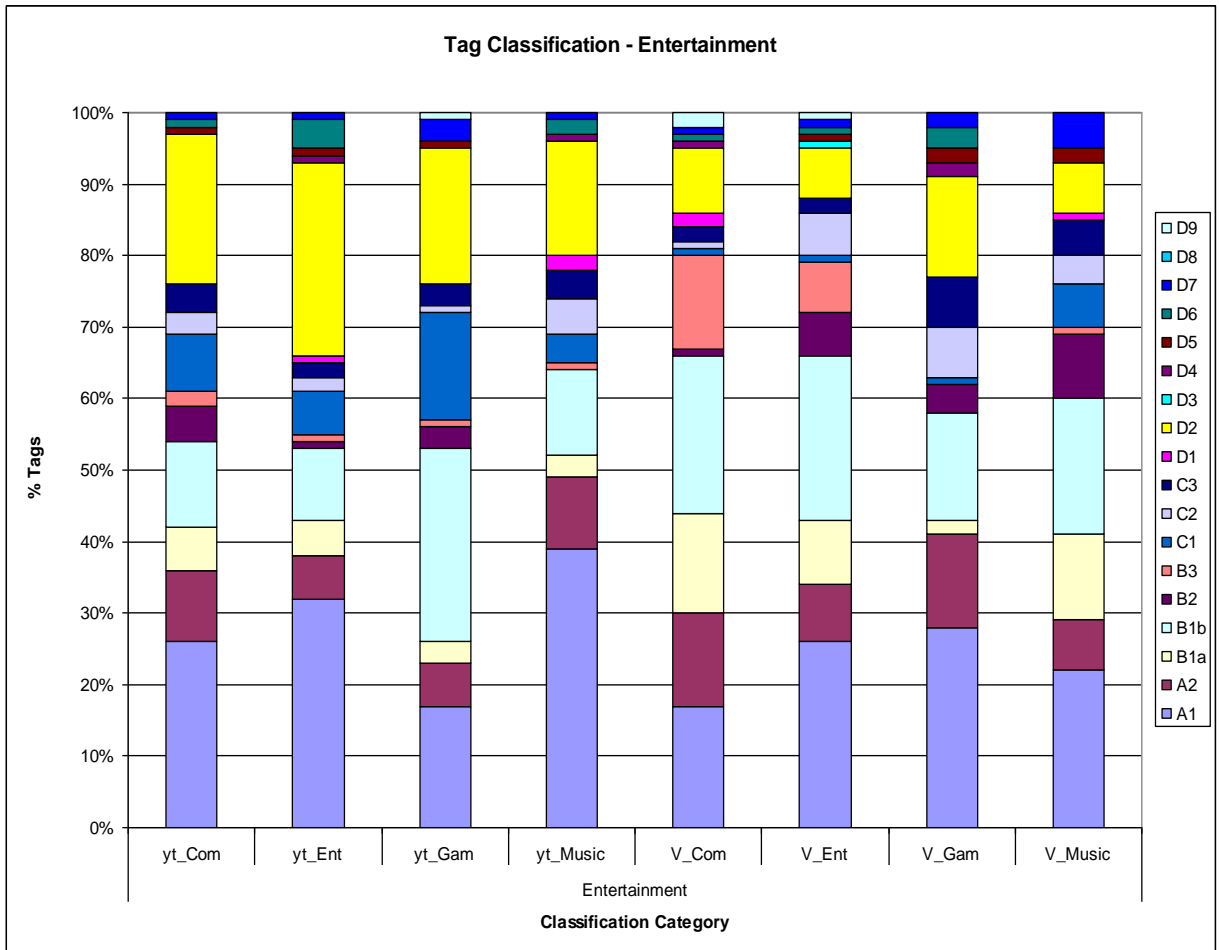


Figure 4-5 The proportions of tag type assigned to entertainment videos.

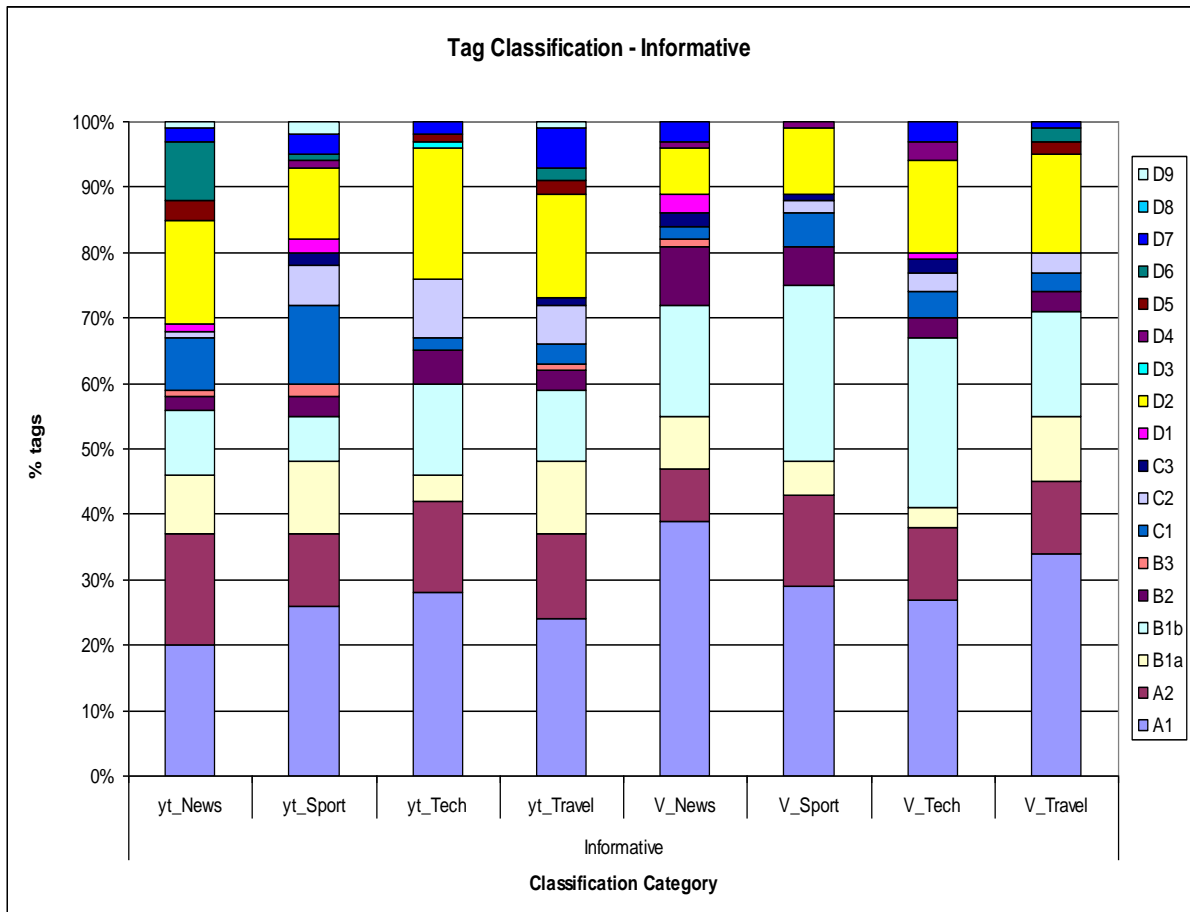


Figure 4-6 The proportions of tag type assigned to informative videos.

Overall 33% of the tags were at the basic level, 26% were specific, 23% were relevant only to an individual or specific group and 18% of tags were non-useful/irrelevant tags. Vidder had more basic descriptions (A1), more specific descriptions (B1a and B1b) and more subjective (B2 and B3) tags than YouTube. YouTube has more social (C) tags with the exception of tags that denote ownership (C3) that were more prevalent on Vidder. YouTube also had more irrelevant (D) tags than Vidder with more Multiword tags (D2) and more Foreign Word tags (D6) in particular (see Table 4-6 for the proportions of each tag type). More specific descriptions of people/objects (B1b), more subjective about (B2) and more subjective opinion expression (B3) tags

were assigned to videos in the entertainment category whereas more basic objective (A1, A2) and more specific descriptions of places/events (B1a) were assigned to informative videos. Informative videos had more foreign word (D6) and more conjunctions and prepositions (D7) tags than entertainment videos. Entertainment videos had more category refining (C1) tags and tags that denote ownership (C3) than informative (see Table 4-7 for the proportions of each tag type). In total, more basic objective (A) tags were assigned to informative videos; more specific objective and subjective (B) tags as well as more social (C) tags were assigned to entertainment videos. There was a similar proportion of irrelevant (D) tags between the two video categories.

Table 4-6 The amount of tags in each tag type classification category for YouTube and Vidler.

Tag Type	YouTube	Vidler
A1	212	222
A2	87	85
B1a	52	63
B1b	103	165
B2	22	41
B3	9	22
C1	58	23
C2	33	26
C3	16	21
D1	6	7
D2	146	83
D3	1	1
D4	3	8
D5	9	7
D6	19	7
D7	19	16
D9	5	3

Table 4-7 The number of tags in each tag type classification category for Entertainment and Informative videos.

Tag Type	Entertainment	Informative
A1	207	227
A2	72	99
B1a	55	61
B1b	140	128
B2	29	34
B3	26	5
C1	42	39
C2	29	30
C3	29	8
D1	6	7
D2	120	109
D3	1	1
D4	5	6
D5	8	8
D6	12	14
D7	15	20
D9	4	4
Total	800	800

Objective language is used more than subjective language for all categories of video in both systems, especially in informative videos. Users of both systems are more likely to describe at a basic level what the video is 'of' than tag to categorise it. Viddler users use more specific objective language to describe video content than YouTube users; in both systems specific language is used to describe people more than places. YouTube has a more international audience than Viddler evident by a large number of non-English tags. More self-reference tags are found in Viddler, this could indicate the presence of social communication, or self-organisation activity within the Viddler community. The only method of video categorisation on Viddler

is through tags, as a result user perception of how useful tags are appears stronger in Viddler with users entering less irrelevant tags than on YouTube and fewer multi-word tags that are less useful for categorisation. As tags are visible and used to order the video library users are more careful about what tags they enter as the tags have more perceived purpose than on YouTube.

Chi square tests were conducted to test for significant differences between the two systems, categories and tag types. A significant difference in tag type between YouTube and Viddler was found $p=.000$. There was no significant difference ($p=.079$) between the amount of objective and subjective tags on Viddler; there was a significant difference between subjective and objective tags on YouTube ($p=.008$). A total of 22 Opinion Expression tags were found on Viddler compared to 9 on YouTube and a Fisher's exact test (used because some expected values were less than 5) was significant at $p=.028$. Viddler users enter more opinion expression tags than YouTube users. The difference between tag type for entertainment and informative videos in both systems combined was significant ($p=.005$). There was no significant difference between the amount of objective and subjective tags in entertainment videos ($p=.160$) but there was a significant difference between the amount of objective and subjective tags assigned to informative videos ($p=.016$). A total of 26 Opinion Expression tags were found in entertainment compared to 5 in informative, Fisher's exact test found a significant difference ($p=.000$). Users express opinion through tags more for entertainment videos than informative videos.

4.3 Conclusion

The weak negative correlation between views and tags indicates that Viddler users rarely engage in collaborative tagging despite the facility being available. This suggests a preference to tag the videos they upload rather than the videos that they view. However, without knowing who assigned a tag it is impossible to confidently prove or disprove collaborative tagging. There is no evidence to suggest that popularity is affected by the amount of tags; a video does not receive more tags because users rate the content higher, nor does it receive more views because a large number of tags has made it more visible. As Viddler categorizes videos only by tag, there is an incentive for the video owner to enter a large amount of tags to increase the amount of categories the video will be assigned to; although the findings suggest a weak tendency for this to deter viewers. There is no such incentive on YouTube and yet the results found a stronger relationship between tags and views on YouTube. This suggests that using tags as extra textual data is useful at improving findability rather than using tags as the sole means of categorization. The lack of empirical evidence of collaborative tagging on Viddler was disappointing as it was not possible to compare tag type between broad and narrow folksonomies. If collaborative tagging exists, why do users not engage in it? Users might need encouragement to collaboratively tag videos (using tagging games in this research).

Users of both systems are more likely to tag a video if it is entertaining rather than informative and use subjective language. There is evidence that users in both systems tag as describers rather than categorisers, despite tags on Viddler being used as the only method to categorise videos. There is evidence that Viddler users give more thought to the tags they enter; there is more irrelevance on YouTube. They are used as additional textual data, but the importance of this is not explained to the users and

they are not actively encouraged to tag. This supports the previous study 4.1 that YouTube users have poor tagging practices and low perception of use for tags. This difference in perceived use had an impact on the amount of tags entered and their quality. More tags are entered per video on Viddler; tags are more relevant to the content and use more specific objective and subjective language.

The purpose of these two preliminary studies was to investigate how users tag videos and find areas where it could be improved. Incentives to tag were higher on Viddler, therefore more tags were generated that were relevant to the video content and described it with a range of tag types. Viddler users rarely engage in collaborative tagging despite tags being their only means of categorisation. Generally, there was a lack of standard practice and no standardised vocabulary. A wider range of language is necessary because users enter more basic objective tags than specific or subjective tags. However, in both systems users describe video rather than categorise it, which is beneficial for increasing textual data to improve video search. This study shows the need for further research into methods to encourage users to tag videos, to improve the perceived usefulness of tags and to encourage users to enter a broader range of type of tag.

N.B. This chapter was published and presented as a poster at Social Networking in Cyberspace Conference 2010: (See Appendix B for the poster)

Greenaway, S. (2010) The Broad Side Of Video Tagging – A Classification Of Tags From YouTube And Viddler. Social Networking in Cyberspace Conference 2010, Wolverhampton.

Greenaway is the author's previous name.

5 The VideoTag Experiment

5.1 Introduction

The video sharing site YouTube relies on text based search methods to index and retrieve videos from its vast library. Tagging creates a valuable source of additional textual data to help this process. However, the preliminary studies in the previous chapter supported the notion that tagging is currently inadequate for providing a solution. Users are not encouraged to tag YouTube videos and so social tagging data is not available. The YouTube user interface does not confront users with the ability to tag successfully so many users are unaware of the ability to assign tags (Velsen and Melenhorst, 2009). This chapter investigates whether users can be encouraged to tag YouTube videos through the gamification of a video tagging system. For a useful system it is essential that the game elements chosen encourage users to enter a variety of tag types, generating many descriptive tags that create semantically rich descriptions for videos. The challenge is to apply an understanding of why people play games and which components encourage users to participate in video tagging games. The biggest challenge for any crowdsourcing project, with or without gamification, is attracting users.

There is no agreed set of attributes for the design of a video tagging game. van Velsen and Melenhorst (2009) provide some requirements for the design of a video tagging system however. The system should display the tagging input mechanism prominently accompanied by a brief explanation of the virtues of tagging and its usefulness for re-finding videos. Indexing and personalised output are a users main motivations to tag videos. Users need to feel there is an indexing purpose to their tagging activity. Non-taggers could be attracted to use a video tagging system by

providing a personal list of most used tags whereby each tag links to a list of videos and videos are recommended based on the user's tags. These requirements were tested in further research; Melenhorst and van Velsen (2010) conducted usability studies of four different video tagging interfaces to measure their impacts on user motivation to tag. Condition 1: Tag Box was the control; it simulated the most familiar tagging environment of text box and tag entry button. Condition 2: ChatBot provided a chat facility whilst the users watched a video; users could chat to each other or a chatbot. Tags were derived from the transcript. Condition 3 mimicked del.icio.us and allowed users to bookmark videos. Condition 4 was a game based system where users rated other users' tags by voting for the best tag. 40 participants were recruited; they rated their experience of each tagging mechanism for seven conditions of perceived usefulness and usability. Users found the game condition more fun to use than the control. Appreciation of video content was affected by the tagging mechanism; the tagging input mechanism should not be too intrusive. The authors advocate a user centred design approach: identify user goals and design for these goals.

Shneiderman (2004) proposes five fun features for interface design: Alluring Metaphors, Compelling Content, Attractive Graphics, Appealing Animations and Satisfying Sounds. These fun features correspond to fiction and juiciness as defined by (Juul, 2009). Compared to elements of game design they are simplistic, emphasising the differences between designing a playful interface and an individual game. To rely on usability or playability heuristics alone to design a video tagging game would result in a poorly designed game and create a shallow, uninteresting experience for users. A GWAP needs to be more than an enjoyable interface. The fun features are useful however, for the design of the VideoTag site; before anyone gets

to a game they are first met with the portal website. This needs to be visually appealing, allowing the user to quickly see the purpose of the site and drawing their attention. If it fails they are unlikely to play a game. It is difficult to create a playful design and game elements are restricted due to the interface required to watch and tag a video. The device to which the game is deployed is also restricted. The game must be a browser based game because handheld devices do not have the screen size to maintain good usability. Users are motivated to use a system if it helps them achieve a goal, so the system is a tool (Malone, 1981). The VideoTag system is a tool to tag videos to improve textual data for videos. VideoTag uses YouTube videos because of the API and embedding functionality. Videos remain hosted on the YouTube server and embedded into the VideoTag system. Tag data is stored in the VideoTag database and assigned to the YouTube video by the YouTube video ID but never made available for use on YouTube. As a result, the external goal of this system is not highly motivating, especially as the tags are not being used in an actual video library so users cannot easily identify benefits from their tagging. The process of tagging a video is routine and boring, therefore applying toy like features by making the process into a game can make the activity more enjoyable (Malone, 1981).

VideoTag encourages the free tagging of videos by users via one of two game based systems or one non-game system embedded within the VideoTag portal. The aim is to encourage language at a conceptual level containing words at varying levels of specificity that provide either objective descriptions or subjective interpretation . The improved textual data collected could be used with automatic methods to improve video search. VideoTag gameplay tries to encourage certain types of vocabulary. The ESP Game encourages basic vocabulary rather than less obvious specific words by only assigning points when users agree. VideoTag will try to avoid this. Jain and

Parkes (2009) suggest that users should be encouraged through gameplay to enter rare words first rather than look for an early match. Golder and Huberman (2006) found that users will enter general tags first, perhaps forcing users to tag for a longer period of time might encourage them to think of tags from different levels creating a broader range of tags. Over time users will select a more diverse set of tags to describe a resource (Chi and Mytkowicz, 2007; Golder and Huberman, 2006). It is proven that matching tags typically have a basic vocabulary and will describe the content at a general level (Goh et al., 2011). Experiments will investigate whether unique tags generated by VideoTag contain a more subordinate and subjective vocabulary.

This chapter discusses the design and implementation process of the development of VideoTag. Two game-based video tagging systems, Golden Tag and Top Tag, and one non-game video tagging system, Simply Tag. An iterative model was followed; three iterations of the design process were conducted to develop the VideoTag system as it is deployed today. Each iteration culminated in an experimental phase. The first iteration was the primary design and development process. Informal observations during this period using prototypes were adapted for the first implementation. A soft launch period of informal user testing was conducted before the system was deployed for the phase one experiment. Informal observations and usage statistics were evaluated, revealing several flaws with the system. The design process was conducted again, culminating in a prototype of the phase two system being tested at a one day promotional event at the University of Wolverhampton. Design decisions from this event were implemented and phase two was deployed, which is its current state. The design decisions and results for each experimental phase will be described over the subsequent sections.

5.2 Primary Design

5.2.1 The VideoTag Website Design

The VideoTag website not only creates a portal for two individual games and one non-game system, but acts as a hub for users to establish a sense of community (albeit on a small scale). It was imperative to create a professional, polished design to attract participants and increase feelings of trust and reputation. Wolfson and Case (2000) state that red creates excitement and encourages users to play and play well. Red arouses a user's interest, but blue sustains play. Red is used for the VideoTag logo and homepage design and blue was used in the game interfaces. A block based design was used, with the sections users were required to interact with most positioned first as the user reads the screen from left to right. Users need to register for an account or login before they can interact with any of the VideoTag features but, key features of the website needed to be visible to spark interest. Figure 5-1 shows the VideoTag phase one homepage. The user's eye should be directed to the three tagging systems, followed by the level thermometer and a selection of the tags that other users have entered into the system. Once logged in, the tags change to a selection of tags that the user has entered into the system. The login and registration panel is at the top of the screen; additional navigation aids for less key areas of the website were placed in the footer. A brief description of VideoTag was used to introduce the system to users with a link provided to find out more if required. Reading was kept to a minimum, concentrating on intuitive design with a limited number of clicks; more information was accessible via in-page and navigation links for users who wanted or needed it.

VideoTag had a social media presence but was not deployed as a Facebook app, instead users were offered the ability to share the site on social media using a series of buttons. A decision was made not to allow users to log in via social network accounts because of perceived security issues users may have in how their personal information would be used. Registration was kept to a minimum with little personal information captured in order to speed up the process and to not deter users concerned about divulging personal information to an unknown organisation. It was made clear during the registration process that their email address was only to be used for validation and not for unsolicited emails. To minimise the amount of spam user accounts added into the database, users were asked to confirm their email address was valid via a confirmation link. Whilst this created extra steps and issues with spam filters, it was unavoidable to protect the dataset and restrict access from spambots.



Figure 5-1 The phase one VideoTag homepage design.

The main rewards system is attached to the VideoTag website rather than individual games. Points earned in the individual games are added to a total site-wide score. Upon login users are informed of their current level as part of the login success notification and once they are at Tea Boy level or higher their username appears on the level thermometer. The Fantasy (Malone, 1981) or Fiction (Juul, 2009) alludes to VideoTag as a TV Company, where progressing through the ranks gives some control over what other users watch. As casual game players are the target audience, the fiction needs to be light hearted, cartoony and offer plenty of juiciness (Juul, 2009) to counteract the mundane activity of tagging videos. Level ranks listed below are based on roles in a TV company:

1. Commissioner
2. Director
3. Producer
4. Researcher
5. Camera Operator
6. Boom Operator
7. Gaffer
8. Best Boy
9. Make-up Artist
10. Dolly Grip
11. Runner
12. Tea boy
13. Intern

All users start out as an Intern. They can quickly progress to Tea Boy normally by playing only one game; then the levels get gradually harder to reach, requiring more time investment. Users can progress through the VideoTag ranks from Tea boy to Commissioner. The challenge for users is to reach the highest levels and to beat other users (perhaps friends) that are also visible on the level thermometer. Users that reach *Researcher* or above can upload videos to VideoTag and they will then get an option to select their own videos to tag before a game or a non-game session. This section is available to all users for information and encouragement via a link in the footer navigation but the actual upload form only appears to registered users at the desired level to avoid spam corruption of the site and dataset. Providing users with the ability to upload videos offers users a sense of control over the system, giving more incentive to invest in the game to tag their own videos and promote their own content to other users. The tags that VideoTag generates will be made available for these users to include as textual data on YouTube.

5.2.2 Golden Tag

5.2.2.1 Synopsis

Golden Tag is a one player browser based game playable on desktop or laptop PCs. A one player format was chosen over the two player match structure developed by Von Ahn and Dabbish (2004) to avoid bias and cheating as suggested by Rafelsberger and Scharl (2009). Barrington *et al.* (2009) found that penalising users if they didn't agree with another player deterred users as they were penalised for their unique opinions. Avoiding the two player match format evades these problems and also allows the investigation of new game formulas that provide variations on the ESP Game template. Golden Tag is built using web technologies using HTML, CSS and JQuery for the front end and PHP to integrate with the MySQL database as the backend. Game elements and graphic design were layered over an essential shell interface of an embedded YouTube video, tag entry form and tag display. Also displayed on screen are a one minute timer, the user's points scored during the game, a space to show feedback for actions and a button to access instructions for 'How To Play' which shows a hidden div that layers over the game rather than navigating the users to a separate page. VideoTag games are first and foremost a video tagging system so the interface was prioritised for functionality. This limited the types of game elements that could be applied. Because the process of tagging involves users thinking of a tag rather than recognising a relevant word it increases the cognitive effort needed to play. Objects must be indentified and content interpreted. This in itself forms part of the game challenge.

The main goal of Golden Tag is to find as many golden tags as possible in one minute, avoiding the pitfalls (tags with high agreement). A golden tag is a tag entered by only one other player and so it is unique. Users cannot score for a golden

tag with a tag that they have entered in the current or in previous games. As more users tag a video, Golden Tag will become easier as there will be more golden tags to find. To counterbalance this and to encourage users to enter specific level tags pitfalls are created. A pitfall is a tag entered by ten or more users; pitfalls are automatically created when a video is tagged enough. As videos are tagged more they will be harder to tag as there will be more pitfalls, but equally there will be more golden tags. This balance between the frustration of finding a pitfall and the joy of finding a golden tag is designed to make the game enjoyable.

The TV company fiction is continued into Golden Tag as users are asked to 'tag through the decades'. Videos are embedded into the interface and masked by a TV screen (see Figure 5-2). In the first level a retro TV typical of the 1950s is used. The theme continues, showing a TV typical of each decade until level 7 - the 2010s when the video is shown playing on a smart phone, and for the final level, a futuristic image is used with the video masked by a pair of glasses. The fiction is not continued into the video content. Whilst this would add novelty to the system, it was not practical when using YouTube videos. However, videos are assigned to an individual level which allows users to unlock new content as they progress through the game. The backgrounds change each time a user progresses to a new level. The original 1950s colour layer was chosen because of its obvious reference to Golden, 1960s changes to red to excite the player that a new level brings variety to the game, reinforced by unlocking new video content; 1970s changes to blue hopefully to incite the user to continue to play. Subsequent levels are a combination of colours all with a blue hue.

Upon selecting to play Golden Tag players are first asked to choose a category of video from one of the following: Comedy, Entertainment, Games, Music, News, Sport, Technology, Travel. They are then presented with a video chosen at random and asked to tag it for one minute. Players are expected to enter keywords that describe the content of the video; they must enter as many as they can before the video ends. Users score points for the tags they enter. Points are allocated in three ways, 100 points are awarded for a tag that has been entered by at least 2 other players, 200 points are lost if a pitfall is entered, 500 points are awarded for finding a golden tag. No points are scored for entering a tag that is not already assigned to the video and points will only be awarded once it has been assigned by another player. This should prevent players from entering tags that are not relevant to the video and reduce cheating. Juiciness is created by adding sounds as feedback to actions as well as writing feedback to the screen; use of sound needs to be limited so as not to detract from the video. A positive 'woo hoo' sound is played if a user finds a golden tag, a negative sound, similar to the sounds played in old arcade games when a life was lost, is played when a pitfall is found. Points are updated on screen when each tag is entered and the tags entered are displayed on screen to help avoid tag repetition. If a tag is not in the English dictionary then the user is alerted by a message in the feedback panel; tags are still entered into the database in order to not lose valuable tag data like names, acronyms and slang or dialect differences. Once a user has finished tagging a video, they will be given the option to tag another video, to tag the same video again or to end the game (see Figure 5-3). At the end of each round of Golden Tag the points scored are added to the user's running total. Golden Tag levels are based on the total points scored from all Golden Tag games. At the end of each game the user's new running total is compared to the threshold needed to progress to a new level. To accompany text feedback written to the screen upon progression, a round of applause is played. The threshold of points needed to

progress to the next level will be less in earlier levels than in later levels encouraging users to continue playing, and create a more difficult challenge in later levels. This is designed to satisfy casual players who want instant gratification, but also to accommodate players who want a more difficult challenge.

For the first user to tag a video Golden Tag will be impossible because all tags entered will be unique. Until more players have tagged the video no golden tags will have been created and users will not be aware of this. All the tags they enter will form potential golden tags for other players. Users only score points for tags with user agreement greater than 2. In the long term this could be solved with a reward structure for being one of the first people to tag a video. In an attempt to overcome this problem for new users, videos in levels one and two were used extensively during the testing period, creating benchmark data for early adopters. As videos are assigned to individual levels, this limits the number of videos presented to users at each level, so players first trying out the game will be able to score more points and quickly progress to the next level, encouraging repeated play. By forcing multiple users to tag the same videos it also improves the challenge of the game, increasing the amounts of golden tags and pitfalls available in each game. The game must feel easy to begin but hard to complete in order to keep players motivated and to sustain play.



Figure 5-2 The Golden Tag in game interface.



Figure 5-3 The Golden Tag end game interface.

5.2.3 Top Tag

5.2.3.1 Synopsis

Whereas Golden Tag encourages users to enter tags of higher specificity to avoid pitfalls and find unique tags, Top Tag encourages users to enter tags with high agreement. Top Tag was also conceptualised as a tool to create recommendations for tags that are the most relevant to the video content and to rank them by agreement. Tags with higher agreement should be more relevant to the video, albeit at a more basic level. These tags are of most use for search as users currently search using basic terms. Top Tag allows for tags with high agreement to be easily identified from the tag dataset. Top Tag encourages tags that complement the tags entered through Golden Tag, hopefully creating a rich variety of semantic language in the tags entered for each video. The game is based on the popular TV game show Family Fortunes, with users being asked to find the five tags most entered by other users for the video. The Golden Tag game engine was reused with adjustments made for gameplay differences. Elements visible on the screen remain the same as Golden Tag with the addition of a gauge showing a user's total points and how many more points they need to progress to the next level. The same videos are used as in Golden Tag.

Reuse of the game engine should make the two games feel similar reducing the learning curve. Differentiation is created by variations in fiction and gameplay between the two games. The Top Tag design prioritises graphical stimulation. The fiction adopts a theme of being bored in a meeting, lecture or school lesson and doodling in a notebook; watching YouTube videos to fill time (see Figure 5-4). The fiction alludes to the type of player Top Tag would like to attract; casual players looking to play short bursts of a compelling game offering instant gratification to

warrant short term time investment. The aim was to accentuate juiciness as defined by Juul (1999) and create a game that was quite easy, less of a challenge than Golden Tag, with appeal coming more from a playful interface. Users are alerted to success in the game by additional writing on the notebook using highlighter pen emphasis and sounds (see Figure 5-5). Top Tag encourages users to enter tags that most other users will have entered; these tags should be easy to think of at most basic level reducing the cognitive cost. Viewers of the TV show Family Fortunes will know that it is not as easy as it first appears to guess the top answers, creating the challenge in Top Tag. The probability of guessing one top tag is quite high, providing instant gratification and engaging the player. The probability of guessing all five top answers is lower, especially given the one minute time limit. Users might therefore feel compelled to tag the video again to continue to try and find the missing answer(s). The balance between boredom created by not finding any top tags or finding them all too easily and frustration at not finding all the top tags, feeling compelled to find the last tag in the list and the joy when it is found should encourage users to continue to play.



Figure 5-4 The Top Tag in game interface.

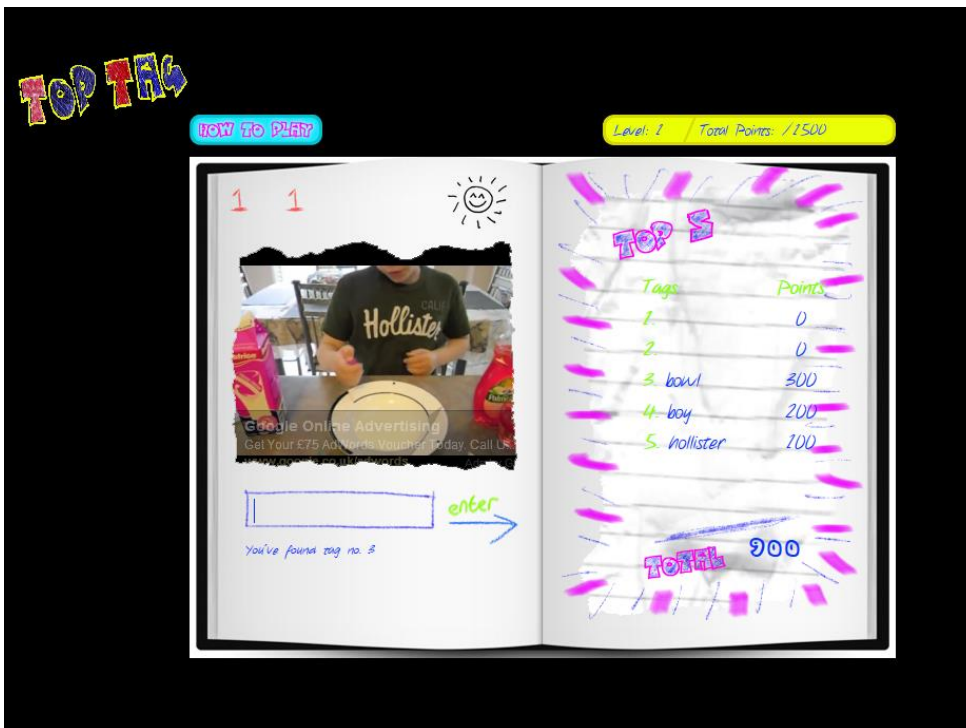


Figure 5-5 The Top Tag interface showing success during a game.

The fundamental gameplay process is similar to Golden Tag. Users are asked to select a category before being presented with a randomly selected video (see Figure 5-6). Users are encouraged to enter as many tags as they can within one minute to try to find the top five tags. Users only score points for tags that appear in the top five and points are scaled depending on the position of the tag in the top five. If they find the top answer they are awarded 500 points, scoring 100 points for each of the other four most popular tags for that video. A cheer is played when a user finds a top tag and the top five area illuminated by highlighter pen marks. At the end of the game if the user has not found all the top tags a sound is played reminiscent of old arcade 'game over' sounds. Users are taken to a game-over screen (see Figure 5-6) where they are told of the total number of points they scored, whether they have progressed to a new level and are given the option to play again or quit. Levels are not as obvious as in Golden Tag because a new level only unlocks new content and the interface does not change. The difficulty of any level is dependent on how many users have tagged the videos. Once users reach higher levels there is more chance that the videos will have been tagged less, making it harder to find tags with high agreement. This creates a cold start problem. If the video has not been tagged yet then there are no popular tags so it is impossible to score points. By adhering to the level structure of videos used in Golden Tag and including tag data from the testing phase, the cold start problem is alleviated in lower levels. If it is encountered it may be seen as an increase in difficulty and part of the challenge. If encountered too early then the game will seem too difficult and deter players, boredom and frustration will become out of balance. In the long term this problem could be solved by using automatic methods to extract benchmark data for the videos albeit limited and representative of low-level features. Top Tag encourages players to enter plenty of tags for the video whilst trying to find the top answers. They are only encouraged to

enter relevant tags. There is no incentive to 'cheat' and enter any tag. Users are alerted if a tag they enter is not in the English dictionary but the tags are still entered into the tag database.



Figure 5-6 The Top Tag select a category and game-over screens.

5.2.4 Simply Tag

In order to evaluate whether games encourage users to tag videos a non game system was added to VideoTag as a control. It was also created to provide users with an opportunity to browse the content of the VideoTag system and to feel that their input mattered to the system. Simply Tag provided users with an alternative tagging method if they were motivated to tag videos but had no interest in playing games. Users could tag videos without engaging in any game elements. Simply Tag avoided some of the limitations of the games if users were interested in the video content. A non-game system is also relevant due to barriers to user enjoyment and output quality found with games as described by Goh *et al.* (2011) that could affect some users:

- Gameplay – competition created by scoring only upon agreement with another player encourages basic level tags as there is more chance of agreement if using general, obvious descriptive terms.
- Time Limit – Placing a time limit on players increases the challenge but also feelings of frustration. Typing speed can be a barrier to the game. It is easier to think of a basic level tag under a time limit. Users spend less time thinking about tags in a game environment.
- Scoring – pressure to score in a fixed amount of time induces frustration that may result in poorer quality tags and reduce user enjoyment.

Simply Tag gives the user greater freedom of choice and allows access to the video library. Users can either browse all videos tagged with a certain tag, or browse all videos in the Golden Tag categories. They are offered the opportunity to upload their own videos, but only if they are at Researcher level or have entered 3000 tags into Simply Tag. For game players attracted to VideoTag, Simply Tag offers a layer of Easy Fun (Lazzaro, 2004) away from the game play to appeal to an explorer player type. The interface is simple and clean. There is no fiction, few graphics and limited feedback; there is little to distract the users from the tagging activity. The strap line 'watch tag explore' summarises Simply Tag. Users are encouraged to explore the videos and the tags in the system. Simply Tag should increase perceived usefulness if users could see how the tags they enter can be used. Users will be motivated to use Simply Tag either by content alone or by an interest in the usefulness of tagging videos. Users need to be convinced of the added value of tagging to participate (Melenhorst and van Velsen, 2010). Indexing and personalised output are a user's main motivations to tag videos and users need to feel that there is an indexing

purpose to their tagging (van Velsen and Melenhorst, 2009). The VideoTag system and both games do not provide an indexing purpose so by allowing users to browse the video library by tag to find video content, Simply Tag provides this purpose.

When selecting Simply Tag as their preferred tagging method, users are first asked to select a category or a tag (see Figure 5-7). They are then sent to a screen that displays all videos for them to browse (see Figure 5-8). No textual descriptions are given, just the title, duration and a thumbnail. Users select a video to tag and progress to the tagging interface. The system uses the Golden Tag engine stripped of all elements of gameplay and controls. The interface is simple including only essential elements, the video, tag entry form and tags entered are displayed. Tags entered are overlaid on top of the video to conserve space and also limit access to the full screen and related videos controls on the embedded YouTube Player to keep users in the VideoTag system. Limited video controls (pause, replay, choose section of video and volume) were made available to increase feelings of control over the system. There is no timer so users can watch the whole video as many times as they like and skip between sections of the video (see Figure 5-9). There is a finish tagging button which ends the Simply Tag session and takes the user to an end page where all tags entered by all users for the video are shown in a tag cloud (see Figure 5-10). Malicious users could tag videos in Simply Tag to get an idea of tags for Top Tag and potentially Golden Tag. However, cheaters will need to be patient because of the random video selection so this should discourage the activity.

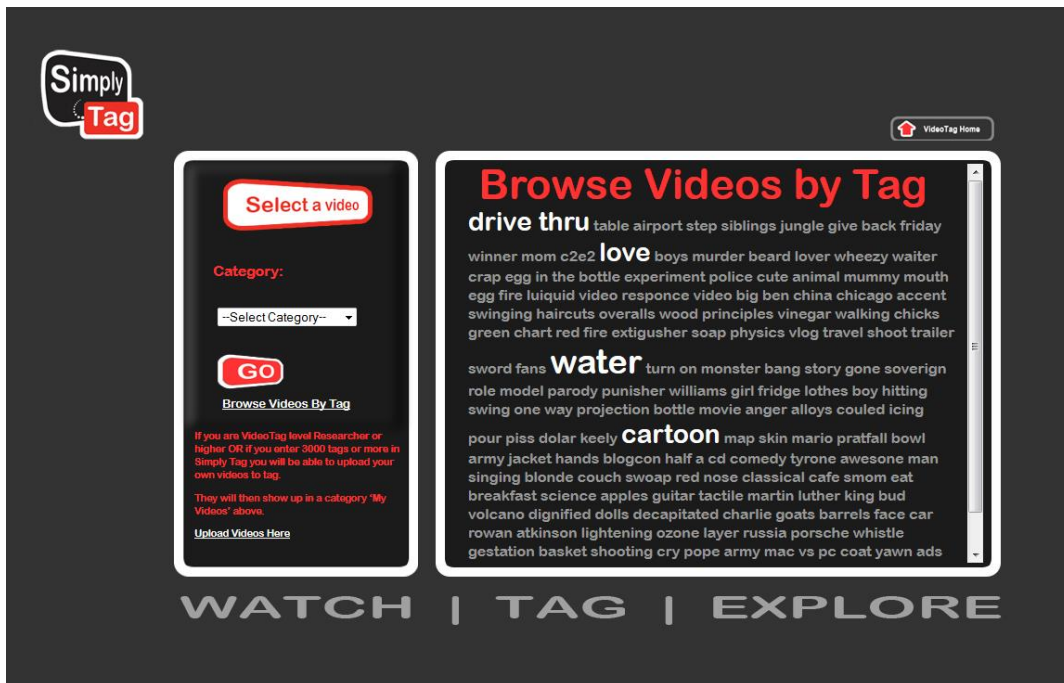


Figure 5-7 The Simply Tag interface to browse by tag.

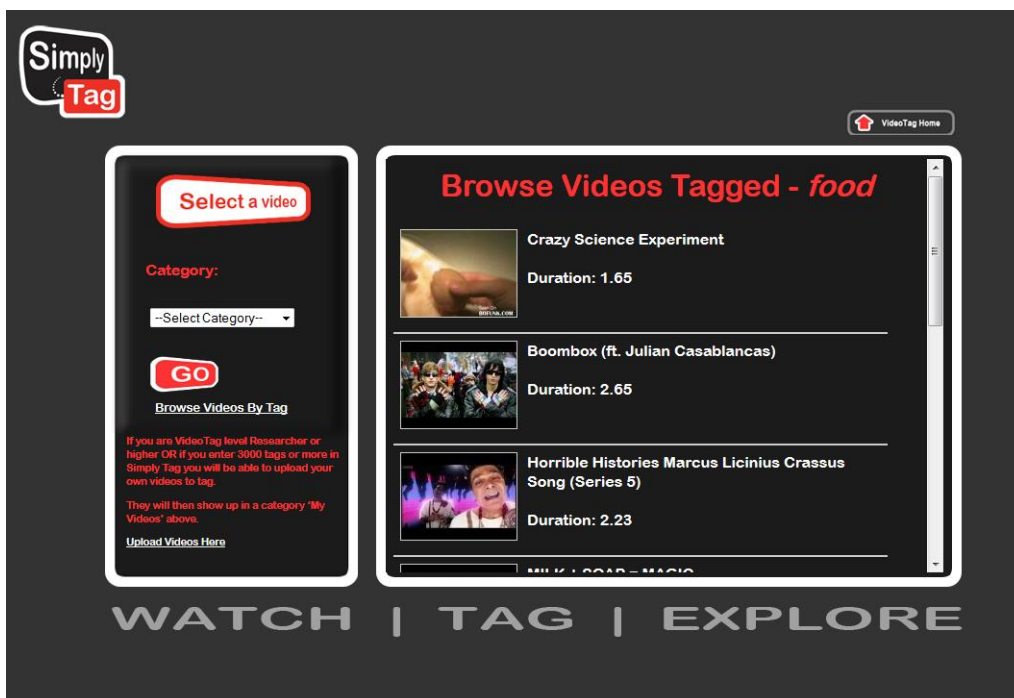


Figure 5-8 The Simply Tag interface to browse for videos.

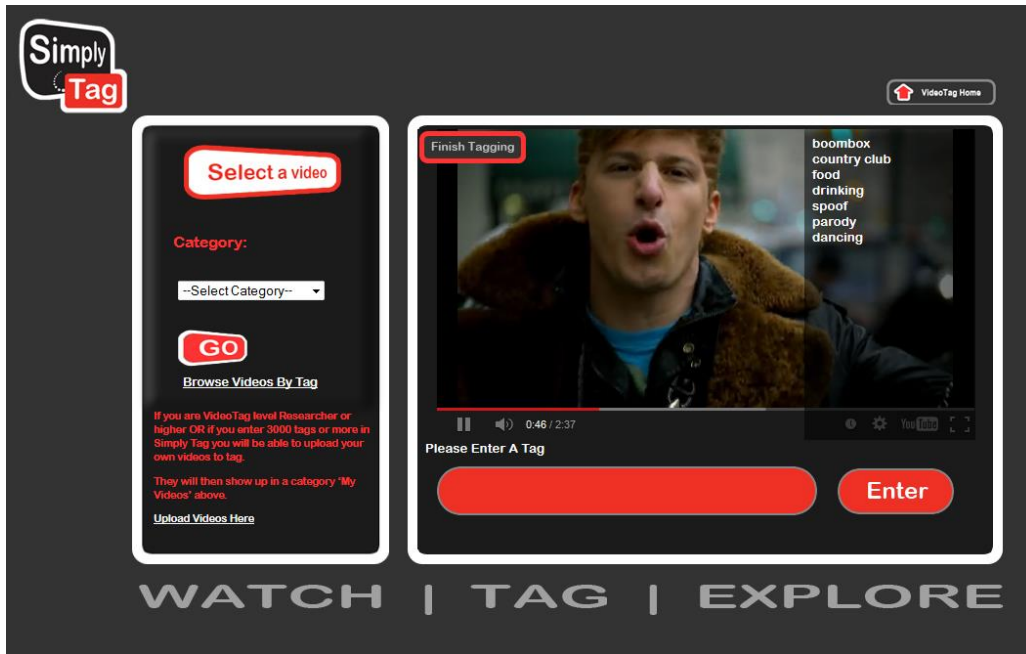


Figure 5-9 The Simply Tag interface to tag a video.

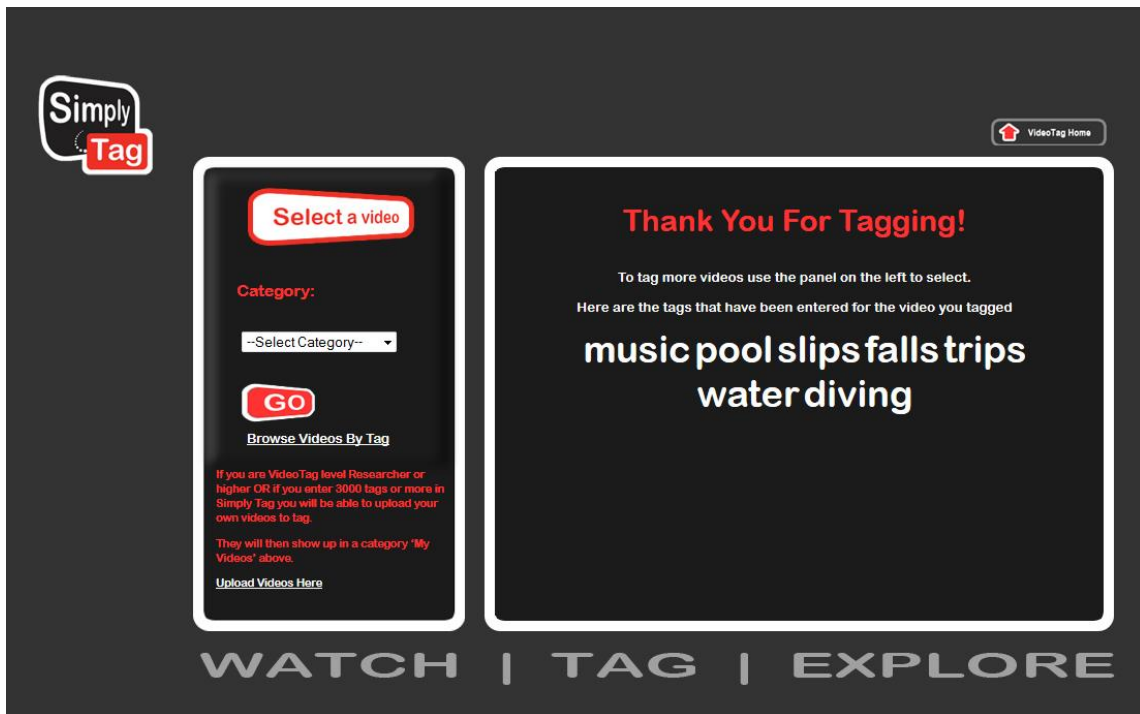


Figure 5-10 The Simply Tag end of session interface.

5.3 Methods

A custom PHP script was used to parse RSS feeds from the YouTube API to obtain a stockpile of YouTube videos. Videos were retrieved on 12th October 2010. Videos were extracted from the eight categories used in the Section 4.2 preliminary study. The RSS feed was ordered by rating, highest rated first. Videos were only retrieved if they were of less than 3 minutes duration. to allow users to get a good idea of what the video was about within one minute and be able to represent this through tags. It also accommodated the fact that users would not invest long periods of time watching and tagging a video that does not fill a specific need. A timer is used in the games to restrict tagging to one minute so all videos will be tagged for the same amount of time unless their duration is less than one minute in length. The duration of the video is important for Simply Tag, where there is no time limit. YouTube

restricts the amount of videos that can be retrieved via its API to a maximum of 500. Less than 50% of the 500 videos retrieved were under the 3 minute threshold. A further search with videos ordered by view count was also conducted. This yielded 519 in the Comedy category, 462 in Entertainment, 221 in Gaming, 105 in Music, 425 in News, 377 in Sport, 305 in Technology and 462 in Travel. As a result of the limit on duration few videos were retrieved for the Music and Gaming categories. A decision was made on 19th Nov 2010 to retrieve more videos for Music and Gaming but to set the limit at 5 minutes. This resulted in 488 videos for the Music category and 345 for Gaming. In December 2012 an extra category, Education, was added to the games, Coursera videos were used in this category. Videos from the Coursera YouTube channel were captured on 20th December 2012. 114 total unique videos were captured these were then assigned randomly to 8 levels, giving 14 videos per level and 2 bonus videos.

The category with the lowest number of videos was Technology with 305 videos; the limit of videos per level was based on this. Hence, 305 videos divided over 8 categories and rounded down to the nearest whole number gave 38 videos in each category per level. This was further rounded down to 35. A random selection of 35 videos from each category was assigned a level ID of 1-8. The remainder of the videos are assigned to a 'bonus' category. View count and rating were used to choose popular videos on YouTube as these stand less chance of being deleted. This is important for this experiment; it would add an unnecessary layer of frustration to the games to be delivered a video that does not play. To avoid this barrier a custom PHP script was created that sends a server call to the YouTube API and records whether the video exists. The database is updated if a video no longer exists, but the video is not deleted because any tag data generated and assigned to that video is still useful

to the experiment and can still be analysed. To maintain the number of videos per level during the experiment, deleted videos are replaced with a video from the bonus category. If the script finds a video no longer exists, users are presented with a link that refreshes the page retrieving a new video.

A prototype of Golden Tag was presented at the 11th Annual International and Interdisciplinary Conference of the Association of Internet Researchers (AoIR) October 21-23, 2010 University of Gothenburg/Chalmers University of Technology, Gothenburg, Sweden. This provided feedback for future developments. Continuous testing and informal evaluations with the GameFlow model of Sweetser and Wyeth (2005) were conducted throughout the design and implementation phase. Results of the final evaluations will be discussed in Chapter 6. Before the phase one experiment was publically launched, a soft launch and further testing period was established for each system. These were staggered over a six week period. For Golden Tag a soft launch period for further testing began 27th February 2013, the period for Top Tag began 13th March 2013 and Simply Tag began 12th April 2013. During this period the game was available on the website but was not publicised. Data from this period is included in the tag classification (Chapter 7). Each soft launch period ran until the start of the phase one experiment. The data generated during the soft launch period was essential to avoid the cold start problem.

5.4 The Phase One Experiment: Publicity to attract users

5.4.1 Methods

The experiment was launched on April 12th 2013 and ran until June 21st 2013. VideoTag was made available online at www.videotag.co.uk and various methods of

promotion conducted to try to attract users. A press release was sent out explaining the VideoTag project and asking for participants. Unfortunately this was not picked up by any news companies or blogs. A request for participation was also made on numerous academic mailing lists. VideoTag was publicised on Twitter and Facebook, a dedicated Facebook page was created and the @videotag2 twitter account used for promotion. During the experiment to try to encourage users to sign up and play a financial incentive was offered to all users who registered. Monetizing the process provides only extrinsic motivation which can deter users with a real interest, curiosity and passion (Cherry, 2012; Van Velsen and Melenhorst, 2009). If not applied subtly it could have a detrimental effect on the quality of the tags entered. Play is a voluntary activity, to add control limits the opportunity for play. It is in a state of play when users will engage most in the games, rather than using the system. A prize draw offering five prizes was created and applied to the registration process. It was hoped this would be the least detrimental to any users intrinsically motivated to play.

5.4.2 Results

5.4.2.1 Usage statistics

Web traffic is measured as hits, pages, visits and unique visitors. Visits record every user who arrives at the VideoTag homepage, no distinction is made for return visits. Unique visitors are visits grouped by IP address, they do not record return visits by a user and give a more accurate representation of how many people have accessed the website. Hits record how many times individual items are accessed on the page (e.g., images, video); page impressions are the amount of whole pages requested. Table 5-1 shows the web traffic statistics for phase one. Although 174 unique visitors were attracted in April and 218 in May, this traffic only created 13 registered users who

played 154 games or Simply Tag sessions. Statistics for June will be discussed in phase two. Users were encouraged to visit, but not to engage in the system, as suggested by the high bounce rate. The lack of registered users indicates that the prize draw did not motivate users. The promotion was not very successful, and failed with all organisations whose reputation would have encouraged visitors (e.g., tech bloggers). The majority of referrers came from the gamification research community and primarily an old link to VideoTag version one (Greenaway, 2007). Social media did not generate much traffic.

Table 5-1 Web traffic statistics for the phase one experiment.

Month	Unique Visitors	Number of visits	Page Impressions	Bounce Rate <30s
April	174	258	13,226	73.2%
May	218	342	7,033	77.7%

5.4.2.2 Tag Frequency

986 tags were entered by 13 unique users in phase one, 484 via Golden Tag, 454 through Top Tag and 48 through Simply Tag. The average number of tags entered by a user was 76, the minimum was 1, the maximum 220. Most tags were entered only once. In total 76% of the tags were unique; the highest level of tag agreement by users was 10 (see Figure 5-11). In the individual systems tag agreement was low with a maximum of 5 in both games and only 3 in Simply Tag. As a result the proportion of unique tags was high with 76.8% generated by Golden Tag, 74.4% by Top Tag and

81.3% through Simply Tag. Figure 5-11 shows the prevalence of high rank, low frequency tags.

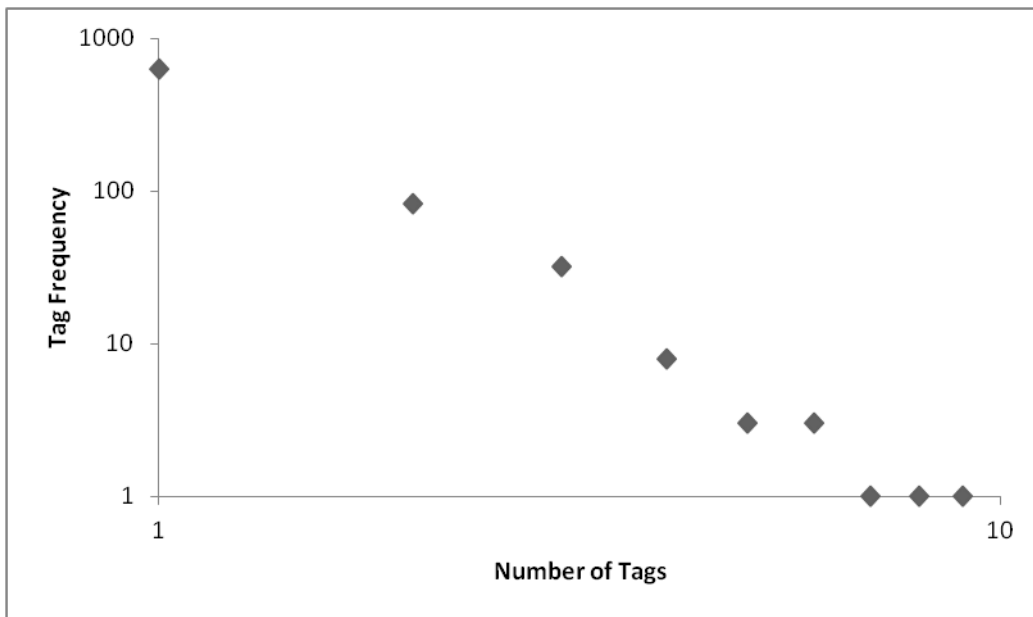


Figure 5-11 Rank tag frequency in phase one (log-log scale).

154 individual games or Simply Tag sessions were recorded. More Golden Tag games (91) than Top Tag games (58) were played. Simply Tag was rarely used with only 5 sessions. Users clearly preferred to tag in a game environment. The average number of tags entered per game was 6, the mode was 3 and the maximum was 21. The most entered tag was *music*.

Table 5-2 shows the ten most frequently entered tags. Users agreed on tags mostly at the basic level except for *funny* which demonstrates user opinions of the video. The presence of a subjective tag in the top ten tags highlights the users' preferences to tag videos that entertained them over those that informed them. 57.9% of the total tags

(see Table 5-3) were entered into categories that entertained, with Comedy and Entertainment being the most frequently chosen categories. Videos were chosen at random once a user had selected a category. 115 videos were tagged. Despite random selection the same videos were selected more than once in different games. Few videos (8) contained only one tag and the modal number of tags per video was 4. The highest number of tags assigned to a video was 44, for a video in the education category. Table 5-3 shows the amount of tags entered per video in phase one. The mean number of tags per video was 9. This is lower than the YouTube (12.42) and Viddler (17.81) means reported in Chapter 4.

Table 5-2 The ten tags with highest user agreement for individual videos in phase one.

Tag	Frequency
music	10
car	8
game	7
dance	6
men	6
song	6
funny	5
girl	5
rap	5

Table 5-3 The amount of videos tagged in each category.

Category	Frequency	Percent
Comedy	223	22.6%
Entertainment	156	15.8%
Gaming	122	12.4%
Education	115	11.7%

Technology	86	8.7%
News	79	8.0%
Sport	73	7.4%
Music	70	7.1%
Travel	62	6.3%

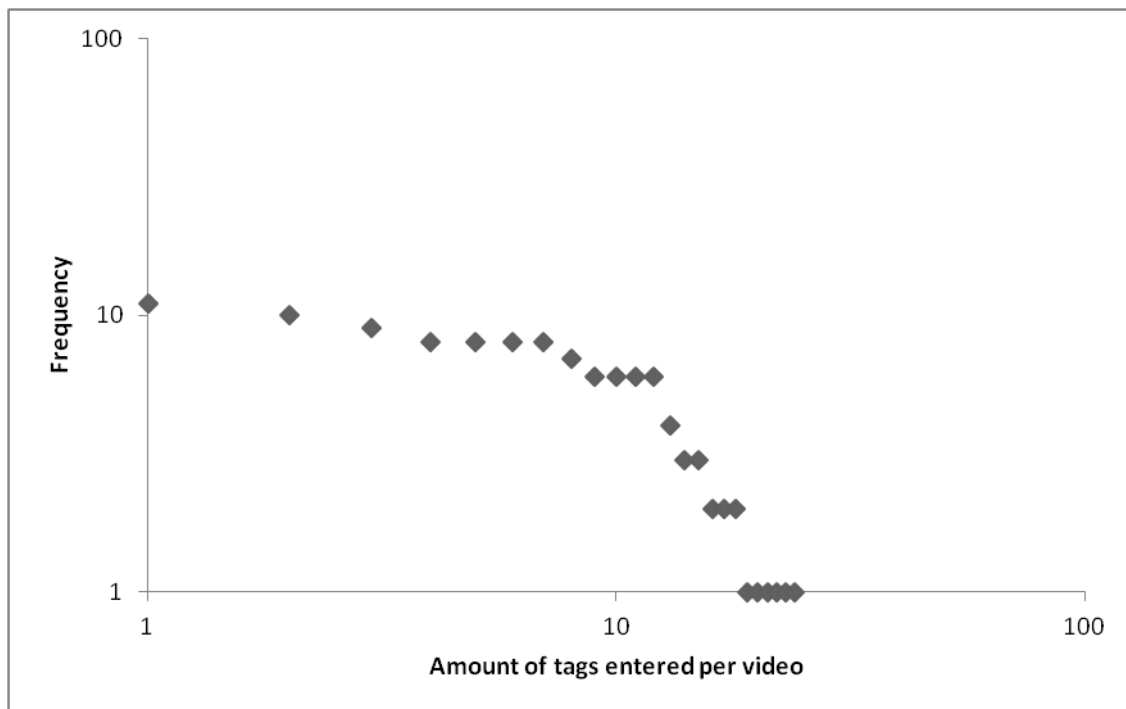


Figure 5-12 Rank frequency of the amount of tags per video in phase one.

5.4.2.3 Observations

No formal user testing was conducted at the end of phase one. Informal user feedback provoked the redesign and implementation of phase two. From this feedback the following observations were made:

- Users attracted through social media often did not know what tagging is.
- Some users did not understand the purpose of the project.
- Users did not want to sign up for an account.

- Some mature users were deterred by a lack of trust and reputation and were reluctant to interact with a system when they did not know who had made it.
- Users did not know what to do or what was expected of them.
- Some users found the process of watching a video, thinking of a tag and typing it in difficult, especially within the time limit.

5.4.3 Discussion

There are cognitive barriers to using VideoTag. The phase one design assumed that users would be familiar with the concept of tagging a video; the level of difficulty was also underestimated. The purpose of the system and descriptions of what to do were not easily accessible; the design was not intuitive for inexperienced internet users. This is a reflection on the type of people that the promotion reached because few tech savvy users or people interested in tagging were attracted. The mass consumption of social media means that users with low computer literacy can use Facebook, but cannot understand a system like VideoTag. For users who did have an understanding and were motivated to 'just try it out' the user account was a barrier. Users with this motivation will not invest a large amount of time in the system, therefore signing up for an account is too time consuming. To sign up for an account involves entering only an email address but this is a time and trust investment in the project. Without being able to see what the system offers before signing up for an account is a deterrent. There are also trust implications. VideoTag was hosted on its own domain and was therefore not instantly recognisable as being affiliated with a trusted or reputable organisation. This was a particular issue for more mature users, who had concerns over how the data they input into VideoTag would be used. This problem was also reported by van Velsen and Melenhorst (2009), who noticed that younger or tech savvy people would try out a video tagging system but middle aged

and less tech savvy individuals were unsure about using it. The issue was less about age than their understanding of social media and technology.

Despite the use of testing data to alleviate some of the cold start problems, the low levels of tag agreement and high levels of unique tags will have made game play difficult in phase one. This will have had a negative impact on user enjoyment and hence on levels of sustained play. The majority of games were played by only a few users. Having a majority of users playing a few games is acceptable in a video tagging system providing that the games attract a large amount of users, which VideoTag phase one did not. The users who registered and participated showed a preference to tag in a game environment rather than Simply Tag, preferring Golden Tag to Top Tag. However, the design of the VideoTag website and the promotion emphasised the games over the non-game system. The preference for Golden Tag may be due to its position on the homepage because users probably read the screen from the top left corner to the bottom right as they would a book. Golden Tag was positioned first on the homepage and was therefore chosen more often. This is most likely the case due to the 'just try it out' motivation rather than being an informed choice.

The VideoTag website design gave little consideration for video content or use of the tags. There was no scope for a sense of community to form. The level thermometer and site wide levels were not well integrated into the site or the games themselves. The design relied on the games themselves being enough of a reason for users to play. Game elements alone are insufficient, video tagging games cannot, for instance, compete with mass appeal games like Candy Crush Saga. More emphasis therefore

needs to be placed on the process of tagging and also the video content. All videos in phase one can be seen on YouTube and found easily through the YouTube categories. With the random selection of videos users have no feeling of control over what they watch. Arends *et al.* (2012) state that personal preference for content affects motivation but random selection of a resource does not. Most participants of crowd sourcing projects have an interest in the activity or the outcome. As the tags are not going to be used by the user and as the user cannot directly see an improvement in their video searches then there is little incentive to use the system. These barriers to use will be addressed in the phase two redesign.

5.5 Phase Two Design

The phase one experiment was disappointing in terms of participants. User opinion gathered informally highlighted several key areas that could be improved. These were to improve perceptions of purpose, improve perceived ease of use for less technically minded users and to transfer the emphasis to content and away from individual games. Rafelsberger and Scharl (2009) suggest that targeting a community of users with specific interests will attract users with a high level of intrinsic motivation to help build a shared knowledge repository. This became the goal for phase two of VideoTag. Both Von Ahn and Dabbish (2004) and Barrington *et al.* (2009) suggest that content affects enjoyment. The redesign concentrated on building a system that was a portal for finding special interest videos. Von Ahn and Dabbish (2004) suggest that creating theme rooms of content could improve user enjoyment in tagging GWAP. The phase two design focussed on VideoTag as a system for curating collections of special interest videos using tags. Tags can be used to recommend a video; if it is good people will tag it, if it is not people won't. Tagging provides a method of filtering content with meaningful descriptions (Shirky, 2005). In turn users recommend videos to other users with similar interests. Users will tag a video that they feel a passion for (Van Velsen and Melenhorst, 2009). Phase one failed to create

a community of users; this was due to a lack of support for community in the system design. Phase two is designed to correct this by offering users a place to curate collections of videos of specific interest to them. The key to phase two was to offer users more choice of content, appealing to niche interests. Special interest groups were targeted to attract people passionate about a specific topic to encourage participation, creating a community of users that would hopefully produce super taggers.

Topics were picked that special interest groups might be interested in. Special interest groups were highlighted through viral trends, internet memes, communities and celebrities on twitter with high followers. A list of possible categories was created, Google Trends was then used to check how popular each category was on YouTube. Checks were made that popular memes were still sufficiently popular and not on a decline, certain categories were chosen because they showed peaks in interest during June and July, these were predominately major music or sporting events in the British calendar. A YouTube search for each category was conducted to check that a sufficient amount of videos existed. Categories were ruled out if they did not fit these criteria. Using a modified version of the custom PHP script used in phase one development, the YouTube API was queried using the VideoTag category title as the query term; results were limited to 15 videos per search. The inadequacy of YouTube search was highlighted during this process, for more specific categories multiple queries were made using different terms. The videos were checked for relevance after each API query and if necessary replacements were found manually. Some categories have more videos than others, uniform categories were not used to give the impression of a community of users.

Table 5-4 lists the categories chosen for phase two and the amount of videos in each category. The ability for users to upload videos was retained. If special interest groups and small communities of users are attracted to VideoTag then they may upload less well known videos to VideoTag or their own content. A barrier to this is not being able to add videos until they have tagged enough videos to reach the Researcher level. The reason for this is that an investment in the project is needed to deter spam. The ability to upload is a reward for investment in the system. To overcome this barrier the opportunity to suggest categories is available to any registered user, or users can Like the Facebook page and suggest categories and videos there, bypassing the level issue, but still allowing for developer moderation of content.

Table 5-4 The number of videos in each phase two category.

Category	No. of Videos
Conspiracy Theories	15
Crazy science experiments	14
Cybermetrics	45
Download Festival 2013	28
Epic Fails	15
Funny Cats	15
Game of Thrones	15
Gamification	17
Glastonbury festival	96
Harlem Shake	15
Historical archive footage	16
Hitler Finds Out	15
Horrible Histories Songs	15
Minecraft	15
One Direction	14
Reverend And The Makers	15

Selena Gomez	15
Sir Alex Ferguson	15
Step Mom Vlogs	33
Stop Motion Animation	15
The 90's	16
The Office	15
Tornados	14
Tour de France	77
Wimbledon Greatest Moments	30
Wolverhampton Civic Hall	15

To further improve the perceived usefulness of the VideoTag system users were given the opportunity to search for videos in VideoTag. Users could search for videos in the VideoTag library from phase one and phase two, the ability to browse videos by tag was also available. This created 'Easy Fun' to appeal to explorer player types (Lazzaro, 2004) as well as an improved sense of use for the tags. A simple search engine was added in phase two that searched for videos only by tag. Information boxes were also placed around the website that gave information about the purpose of the game (see Figure 5-13). Questions such as why tag, why participate and how are tags useful were answered. 80% of people in van Velsen and Melenhorst (2009) did not know what tagging is and this could be replicated in users attracted to VideoTag. Informal user feedback also suggested that this was a problem. Explaining what tagging is and why it is useful using short text notes in graphical information boxes might reduce this barrier to use.

To improve ease of use and further overcome the barrier for less experienced internet users the phase two design added step by step instructions on how to get started. It was graphical to avoid large amounts of text (see Figure 5-13). Block design was used again, each block containing an action that users had to take. The blocks created a

flow chart until users reached the stage of watching and tagging a video. The problem with this is that the content could be seen as too busy and overwhelming for some users. It also requires more clicks from the user before they start tagging, which could be a barrier for more experienced web users. The number of decisions a user has to make before getting to watch and tag also increased, this adds to the cognitive cost for users and could be a deterrent for pick up and play users or users motivated to 'just try it out'. However, the navigation was designed to be easy to follow (see Figure 5-13). Step one directs the user to sign up for an account or login; step two asks them to select a category. The further steps were detailed graphically on the homepage so users could see at a glance what the whole system entailed. However, once a category was selected users were navigated to a page where they could select a video (see Figure 5-14). A link to return to the previous page was available so users could correct their actions. They were also given the option to select a different category. Once a video was selected users were navigated to another page to select a tagging method and they were given the graphical options of Golden Tag, Top Tag or Simply Tag. Again users were given the opportunity to correct their actions. Once a system had been selected the games were identical to phase one with the exception of the category selection page being removed. Few changes were made to the Simply Tag interface; the browse functionality in the left panel was replaced with navigation buttons (Figure 5-16). The end game pages of both games were altered, improving the integration of the level thermometer. Users were also given the opportunity to tag the video they had just tagged again (see Figure 5-15). This was an important alteration in Top Tag to remove the frustration of not being able to try to find the missing top tag and thus to incentivise users to continue to play.

Giving users more control over the system and reducing the amount of random selection could encourage more malicious users. The visibility of tags in the system provided malicious users with the ability to cheat. The removal of the random video selection made the games more vulnerable. However, the tag clouds were not accurate in their display of tag frequency, so if a user wanted to tag a video in Top Tag thinking that they knew the top answers, they could still struggle to find them all. If a user wanted to go to this much trouble to cheat the system then during the process they would still enter beneficial tag data. The login requirement was retained as a deterrent to this because users committed to the project enough to register are less likely to enter malicious data (Trant, 2009).



Figure 5-13 The homepage of the phase two version of VideoTag.



Figure 5-14 The select a video page in phase two.



Figure 5-15 The phase two game over page.



Figure 5-16 Phase two changes to Simply Tag, navigation buttons replace the select video panel.

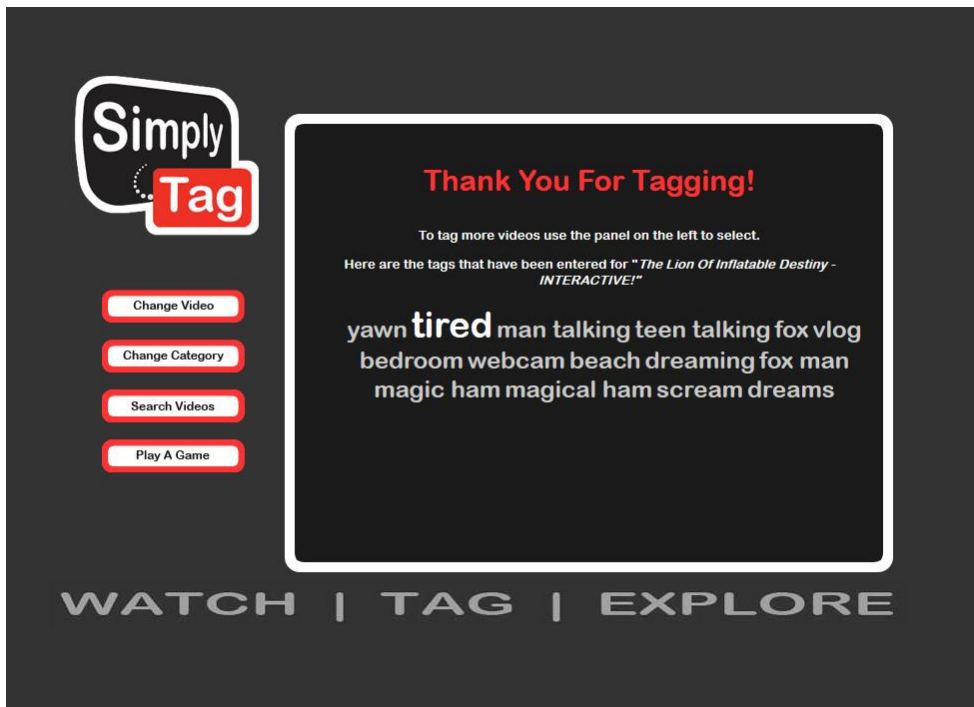


Figure 5-17 Phase two changes to the finish tagging page in Simply Tag.

5.6 Phase Two Prototype – SciFest Experiment

5.6.1 Methods

SciFest was a one day event held at University of Wolverhampton on 22nd June 2013. VideoTag was available as one of the displays to showcase scientific study and research to children. The SciFest version of VideoTag was a restricted version and the user account 'scifest' was created that restricted users to 14 videos within the 'Crazy science experiments' category. The single user account was created to remove the sign up barrier and to encourage players to try it out. Participants were able to see the other categories but could not tag the videos. The aim of showing the other categories to students was to encourage participants to register after the event and continue to play.

5.6.2 Results

Visitors to SciFest played a total of 431 games, entering 2,210 tags for 14 videos. 139 Golden Tag games (32.3%) were played generating 609 tags (27.5%), 260 Top Tag games (60.3%) generated 1211 tags (54.8%) and 32 Simply Tag sessions (7.4%) generated 390 tags (17.6%). The majority of users played just one game, preferring Top Tag. All videos were tagged at least once in all three tagging methods, with the exception of one video that was not tagged in Golden Tag (see Figure 5-18). Table 5-5 shows the mean, mode and maximum number of tags entered in each game and Simply Tag session during SciFest. There was a tendency to enter more tags per session in Simply Tag than games as shown by the increase in mean (10), mode (17) and the maximum (25). Despite being used less and having generated less tags overall, users showed a tendency to enter more tags during a Simply Tag session than during a game. This is probably a result of the time restriction applied to games.

Table 5-5 The mean, mode and maximum tags entered in each game or Simply Tag session.

	Golden Tag	Top Tag	Simply Tag
Mean tags per game	4	5	10
Mode tags per game	4	3	17
Maximum tags per game	13	15	25

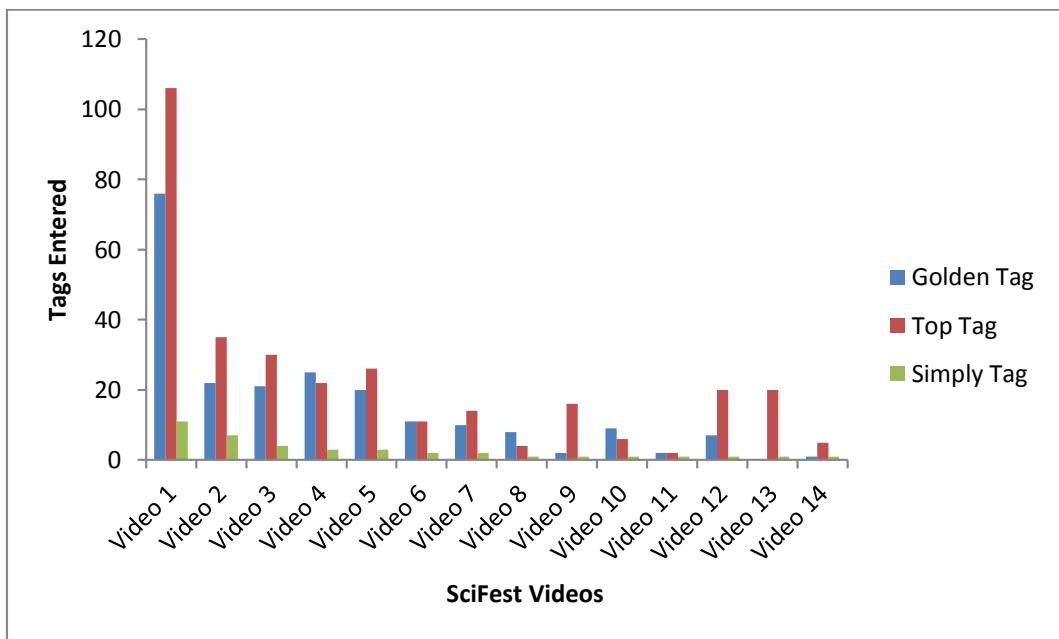


Figure 5-18 Amount of tags per video, per tagging method, in Crazy Science Experiments category (see Table 5-6 for the titles of each of the 14 videos)

Considering the three tagging methods together, the mean number of tags entered per video was 138. This is more than the averages for YouTube (12.42) and Viddler (17.81) (see Chapter 4) and more than phase one (9), but this is due to the small

number of videos that users could choose from. The maximum number of tags entered for a video was 663, the minimum was 33. Table 5-6 shows the proportion of tags assigned to each video during the SciFest experiment. The most tagged video was the first video in the list, Video 1 - *Fire Hands! fire experiments*. The video was tagged 76 times using Golden Tag, 106 times using Top Tag and 11 times using Simply Tag. The second most tagged video, *MILK + SOAP = MAGIC* tagged 22 times using Golden Tag, 35 using Top Tag and 7 using Simply Tag. However, the list changed as it is ordered by most tagged. The video that was first at the start of the day, *Crazy Science Experiment* was tagged third most frequently (4th in Golden Tag). Users during SciFest had no specific interest in the content and would predominantly choose the first video in the list, which was a problem in the phase two design. The original idea was to encourage users to tag their favourite video and have the list ordered by most tagged or ranked by most popular. However if users typically select the first video it is more useful to order videos by random selection.

Table 5-6 The number of tags assigned to each of the 14 videos used in SciFest.

Video	Title	Number of times video was tagged			
		Golden Tag	Top Tag	Simply Tag	Total
Video 1	Fire Hands! fire experiments	76	106	11	193
Video 2	MILK + SOAP = MAGIC	22	35	7	64
Video 3	Crazy Science Experiment	25	22	3	50
Video 4	Crazy Balloons - Easy Science Experiments for Kids	20	26	3	49
Video 5	Crazy science experiment - cool physics demonstration - anti-gravity water glass	21	30	4	55
Video 6	Making Hot Ice (Crazy Science Experiment)	8	4	1	13
Video 7	"Mad Science" Crazy and Dangerous Home Experiments	2	16	1	19

Video 8	Awesome CD Trick - Crazy Science Experiment	9	6	1	16
Video 9	10 Amazing Science Magic Tricks for Parties Part 1	10	14	2	26
Video 10	Crazy Beer Bottle Trick! Bet You Will Always Win!	0	20	1	21
Video 11	Crazy Neodymium Magnet Experiments	1	5	1	7
Video 12	egg in the bottle Science Experiment	11	11	2	24
Video 13	How to make a light bulb with pencil lead. Crazy easy science project	2	2	1	5
Video 14	5 Amazing Science Experiments Using Plastic Bottle Part 4	7	20	1	28

Table 5-7 Ten most entered tags in each system during SciFest.

Top Tag			Golden Tag			Simply Tag		
Tag	Frequency	Percent of total	Tag	Frequency	Percent of total	Tag	Frequency	Percent of total
fire	79	6.5	fire	17	2.8	fire	21	5.4
water	72	5.9	water	13	2.1	bottle	18	4.6
man	55	4.5	egg	13	2.1	water	18	4.6
bottle	51	4.2	glasses	10	1.6	man	15	3.8
glasses	31	2.6	soap	10	1.6	bowl	12	3.1
glass	30	2.5	bottle	9	1.5	liquid	11	2.8
liquid	30	2.5	milk	9	1.5	egg	8	2.1
egg	25	2.1	experiment	8	1.3	soap	8	2.1
bowl	24	2	bowl	6	1	glasses	7	1.8
milk	21	1.7	boy	6	1	hollister	7	1.8

Table 5-7 displays the ten most frequently entered tags during SciFest using the three systems. Top Tag generated a higher proportion of high frequency tags than Golden Tag i.e. the tag 'fire' was entered 79 times in Top Tag compared to 17 in Golden Tag. This suggests that gameplay might have had an effect on tagging behaviour. As Top Tag encourages the user to find the tags entered most frequently there should be

more agreement than in Golden Tag. Moreover, more basic level tags should also be entered into Top Tag. Since users try to find tags that only one other player has entered in Golden Tag, a more specific level would be expected and less agreement. Table 5-7 shows clear differences in the frequency of tags between the three systems, with Golden Tag having the lowest frequency levels. However, the tags themselves show high levels of similarity and are at a basic level. High levels of agreement are unlikely for specific level tags; Figure 5-19 shows the distribution of tag frequency for each system. The prevalence of high rank, low frequency tags in each dataset is clear because 74.2% of tags entered into Golden Tag are unique, compared to 68.3% in Top Tag and 68.9% in Simply Tag. Whilst the numbers of unique tags are still high in Top Tag, a larger proportion of low rank high frequency tags are evident. Fewer tags were entered into Simply Tag and tags that were entered were less likely to appear again and had greater specificity. The differences in distribution suggest that gameplay affected the tags users entered. The larger proportion of unique tags harvested by Golden Tag implies that the gameplay of Golden Tag encouraged users to enter more unique tags when trying to find tags entered by only one other user.

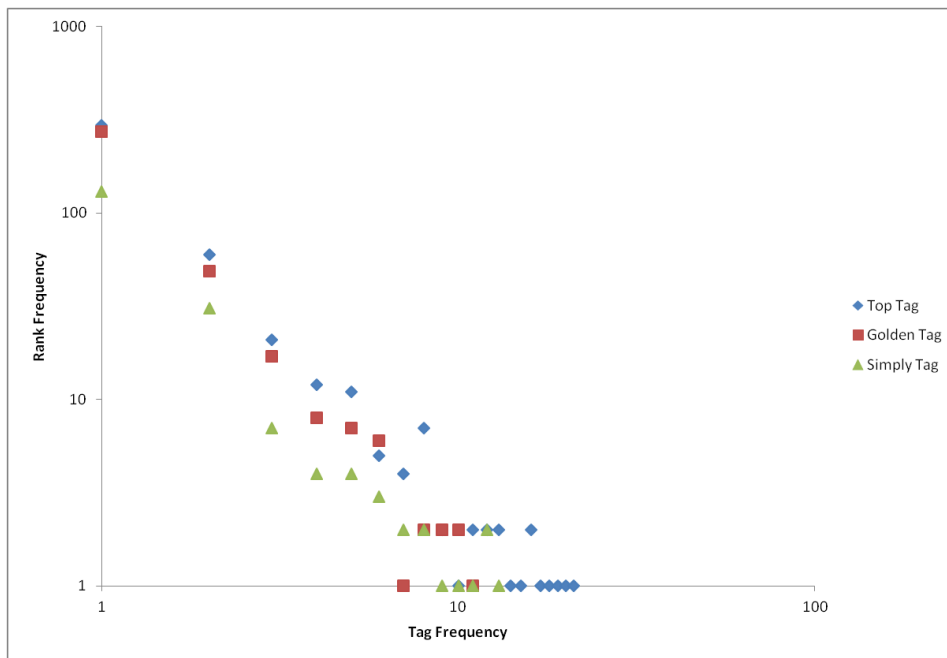


Figure 5-19 Rank frequency of tags entered into the three tagging systems during SciFest.

Golden Tag was considered to be too hard with SciFest users stating that it was impossible to find Golden Tags even though overall 66.9% of tags were unique. Taking the most tagged video *Fire Hands! fire experiments* as an example; there are 121 Golden Tags for this video and yet players found it incredibly hard to find a Golden Tag. Removing the typos and spelling mistakes left 83 genuine Golden Tags.

5.6.3 Design Decisions Based On Analysis of the SciFest Prototype

Informal observations of the SciFest prototype revealed that the majority of users preferred Top Tag to Golden Tag; finding it easier to score points in Top Tag with some users finding it difficult to find Golden Tags. Users preferred to play a game rather than Simply Tag. As was highlighted in phase one, some users found the cognitive effort of watching a video and entering tags in the time limit difficult.

Despite efforts to make how to play and the purpose of VideoTag more accessible and clearer, some users still needed verbal explanations of the system. Design decisions based on the results of the SciFest experiment and informal observations were applied to the VideoTag website before further promotion for the continuation of the phase two experiment. Videos were ordered at random rather than by most tagged as the majority of users selected the first video on the list rather than making a choice based on interest in the content. Ordering videos by random should give more balanced weight to the amount of tags per video and with enough users, allow for any genuine favourites to be highlighted. The positions of Top Tag and Golden Tag were changed, moving Top Tag to the first position reading left to right across the screen (see Figure 5-20). Top Tag is the preferred game, it was played considerably more and users commented that it was easier to play. Game testers also commented that Top Tag is more fun to play and gives more instant gratification. Therefore if users are motivated to 'just try it out' and play one game it would be best for them to play the more enjoyable game.



Figure 5-20 Selecting a tagging system in phase two.

The user account is a barrier to play for less motivated users but the usage statistics showed that it was imperative to avoid infiltration by robots and spam tag data. A guest account was introduced for phase two, but guest users will have restricted access and game elements will be limited. The aim of this is to attract more users and perhaps some will convert from guests to registered users. Some users were frustrated by the one minute timer and wished that they could watch the end of the video and tag for longer. However, users already have three choices to make before watching and tagging a video and for some it was observed that this choice is a barrier to play. Many users (parents and children) needed an explanation of the purpose of VideoTag. One parent at SciFest asked “What has this got to do with

science?" Feedback from phase one also highlighted this issue. This is a major barrier to use; if users have no sense of purpose to their actions they will not interact with the system. Equally, potential players need clear, well defined goals and need to feel the game is easy to learn. Phase two offers more information and simplified instructions, but if users do not to read them then it is impossible to overcome this barrier. Learning to play should be part of the fun, incorporated into the initial versions of the game. This is difficult because VideoTag is primarily a video tagging system and not a game, but it should be considered for future video tagging games.

5.6.4 Summary

The user behaviour observed during SciFest offered explanations to the disappointing uptake during the phase one experiment. Visitors to SciFest expressed an interest in the concept of the research and were enthusiastic about playing a game and scoring points. Unfortunately, most participants left after playing one game. Some users found VideoTag too difficult and were not motivated to persevere and conquer the challenges. Engagement is affected because gameplay does not effectively balance frustration and boredom. This could be because the process of watching a video and tagging the content is too high a cognitive cost that it becomes a barrier to engagement. Users enjoyed the content but for some, the process of tagging and the game elements interrupted their enjoyment of the video which in turn drove them away. It was observed that participants had little comprehension of what tagging is and why it is useful, this needed explaining before they would interact with the system. Promoting VideoTag as a set of games had limited success, for phase two focus of the promotion will concentrate on the video content.

5.7 Phase Two Experiment

5.7.1 Methods

As with phase one, users were attracted to use VideoTag through promotion via press release and social media. The financial incentive to register for an account was retained. Phase two strived to create an improved perception of purpose and to attract users interested more in content than playing games. The phase one website redesign with the addition of highly specific video categories allowed for promotion to special interest groups. Video categories were chosen based on current events at the time of the experiment: current popular video viral trends and memes. In order to alleviate some of the trust issues highlighted in phase one, various methods were conducted to bolster the reputation of VideoTag through promotion by trusted organisations and high profile Twitter users. A press release was sent out to various technology blogs, posts were made on related forums and two Wolverhampton based organisations with high social media profiles were contacted, but this generated little interest.

Extensive promotion was conducted by the author between June 23rd and July 1st using Facebook and Twitter. Promotion on Twitter used hashtags to highlight video content related to specific current events (e.g., Wimbledon, Tour de France, Glastonbury and Download Festivals) or to target fans of bands playing at the festivals (e.g., Slipknot, Stone Sour). Niche groups were identified on Twitter and the member with most influence was contacted. The most notable was a support group accessed via the hashtag '*#twitterstepmoms*'. This predominantly American group has many bloggers and prolific tweeters, the most high profile of which, @Cafesmom was contacted and agreed to promote VideoTag to her followers and to publicise the experiment on her blog.

5.7.2 Results

5.7.2.1 Usage Statistics

Table 5-8 displays the monthly web traffic to the VideoTag website during the phase two experiment. June contains traffic for phase one as well as the SciFest prototype. SciFest is the cause of the large increase in pages and hits during June, other spikes in activity correlate to publicity and user studies. These statistics indicate that although publicity attracted more visitors to the site than in phase one, again few visitors were turned into registered users. Of the 1,795 unique visitors recorded over the 6 month period, 12 were registered users, 4 were return users who registered during phase one and some used the guest account. This and the high bounce rate highlight the tendency for visitors to leave VideoTag without playing a game, only 124 games were played. The majority of users left within 30 seconds and only visited the homepage. Inspection of referral links revealed that many visits were by robots, in addition over 2000 spam user accounts were created. These spam accounts were captured and isolated from the useful data by adding in a confirmation email link when users sign up. It is for this reason the user account was created making games only accessible to registered users. These precautions were created to prevent spam tag data being entered through game or simply tag interfaces.

Whilst registering for an account is a barrier to the 'just try it out' motivation, registered users will have less reason to enter malicious tag data (Trant, 2006). Excluding traffic by spambots, most traffic came through Twitter links and a direct link from the Cafesmom blog. Users that were attracted by the twitterstepmoms promotion may not have created many registered accounts, but the step mom vlogs were highly tagged in Simply Tag by the guest account. The fact that traffic is higher than actual users, coupled with high bounce rates indicates that users that were

attracted to the site chose not to register or play. The Cafesmom promotion proves that users with a specific interest can be encouraged to participate, but an interest in the content alone is not enough of a reason to use VideoTag. Users were perhaps not inspired by the gameplay or saw little purpose in the system or tagging the videos. Intensive promotion on social media did not generate much traffic and neither did the press release. The Cafesmom experiment proved that links from a trusted source with reputation can increase users. Unfortunately, the promotion did not seem to generate many users interested in tagging videos.

Table 5-8 Web traffic during the phase two experiment (N.B., June also includes phase one and SciFest traffic).

Month (2013)	Unique Visitors	Number of visits	Page Impressions	Bounce Rate <30s
June	236	515	65,601	73.30%
July	386	620	11,673	74.50%
August	186	316	1,921	81%
September	232	484	7,058	68.10%
October	197	654	9,204	68.50%
November	289	1079	11,523	67.70%
December	269	855	5,235	81.90%

5.7.2.2 Tag Frequency

124 games were played in phase two by 12 unique users and the guest account. Users played 47 games of Golden Tag, 42 of Top Tag and 35 sessions of Simply Tag. Table 5-9 shows the number of tags entered using each system. In total 1018 tags were entered: 305 tags in Golden Tag, 301 in Top Tag and 412 in Simply Tag. The mean number of tags entered per game (or Simply Tag session) was 8, the mode 6, the maximum 34. For Simply Tag the mean number of tags per session was 12 compared to 6 for Golden Tag and 7 for Top Tag. Most users entered <10 tags per game/session.

The guest account generated 343 tags, 297 via Simply Tag, 35 via Top Tag, 9 via Golden Tag. Two games of Golden Tag were played, one round of Top Tag and multiple sessions were recorded in Simply Tag. This data suggests that by emphasising video content phase two users chose to tag the whole video rather than play a game. However, this is only true for one video category. The Simply Tag sessions coincided with promotion on Twitter to a special interest group ‘twitterstepmoms’. Table 5-10 shows ‘Step Mom Vlogs’ was the most tagged category in phase two. This promotion created 4 registered users but, only 2 tagged a video.

Table 5-9 Number of tags entered using each system during the phase two experiment.

	Golden Tag	Top Tag	Simply Tag
Total tags per system	305	301	412
Tags entered using guest account	9	35	297
Mean tags per system	6	7	12

Table 5-11 highlights the differences in tags between the three systems. Tags in Simply Tag relate to the Step Mom Vlogs videos whereas they are not present in the most frequent tags for Golden Tag and Top Tag. Tag frequency is lower than in phase one. This could indicate that more tags of higher specificity were assigned to phase two videos. However, there was also a lower proportion of unique tags generated during phase two: 65.6% of tags were unique in Golden Tag, 69.4% in Top Tag and 73.3% in Simply Tag. These findings indicate that in phase two individual gameplay of Golden Tag and Top Tag did not affect the tags that users entered. More unique tags were entered into Top Tag than Golden Tag; if gameplay was affecting tagging behaviour then the reverse would be true. However, Table 5-11 indicates that users did enter more specific level tags into Golden Tag. Perhaps the specificity of

content had more effect on the tags users entered than the system used. This will be investigated further through tag classifications in Chapter 7.

Table 5-10 The most tagged categories in phase two.

Category	Frequency	Percent
Step Mom Vlogs	298	29.3
Glastonbury Festival	145	14.3
The Office	106	10.4
Crazy Science Experiments	101	9.9
Wolverhampton Civic Hall	80	7.8
Horrible Histories Songs	67	6.6
Download Festival 2013	49	4.8
Hitler Finds Out	35	3.4
Wimbledon Greatest Moments	34	3.3
Funny Cats	33	3.2
Stop Motion Animation	18	1.8
Conspiracy Theories	15	1.5
Epic Fails	14	1.4
Reverend And The Makers	9	0.9
Game Of Thrones	8	0.8
Harlem Shake	5	0.5
The 90s	1	0.1

Table 5-11 The ten most frequently entered tags in each system phase two.

Golden Tag		Top Tag		Simply Tag	
Tag	Freq	Tag	Freq	Tag	Freq
glasses	5	tents	6	smom	9
guitar	5	driving	4	advice	6
drums	4	glasses	4	kids	6
festival	4	man	4	step mom	5
live	4	bins	3	support	5
office	4	g	3	vlog	5
dwight	3	live	3	behaviour	4

funny	3	map	3	help	3
game	3	men	3	love	3
hair	3	music	3	step mothers	3

Despite 18 users registering during the six month period of the phase two experiment, only 12 tagged a video. The mean number of tags entered per user was 56, the minimum was 1 and the maximum 206. 84 videos were tagged in total, the mean number of tags per video was 12, the mode was 6, the maximum 62. Although the mean is lower than in phase one (19) the mode (4) and maximum (44) were higher in phase two (see Table 5-12). The majority of videos were assigned more than 10 tags. Despite a random ordering of the categories users still chose similar videos to tag (see Figure 5-21). The most tagged video was from the Glastonbury Festival category, which was the second most tagged category (see Table 5-10). Even though videos were selected to appeal to specific groups of users (e.g. Cybermetrics videos for the Statistical Cybermetrics group members at University of Wolverhampton or Gamification to attract members of the gamification research forum), many of these videos were not tagged. The results of phase two support the finding from phase one that users prefer to tag videos that entertain them, especially in a game environment.

Table 5-12 A comparison of the mean, mode and maximum number of tags entered per video during phase one and phase two.

	Phase one	Phase two
Mean tags per video	19	12
Mode tags per video	4	6
Maximum tags per video	44	62

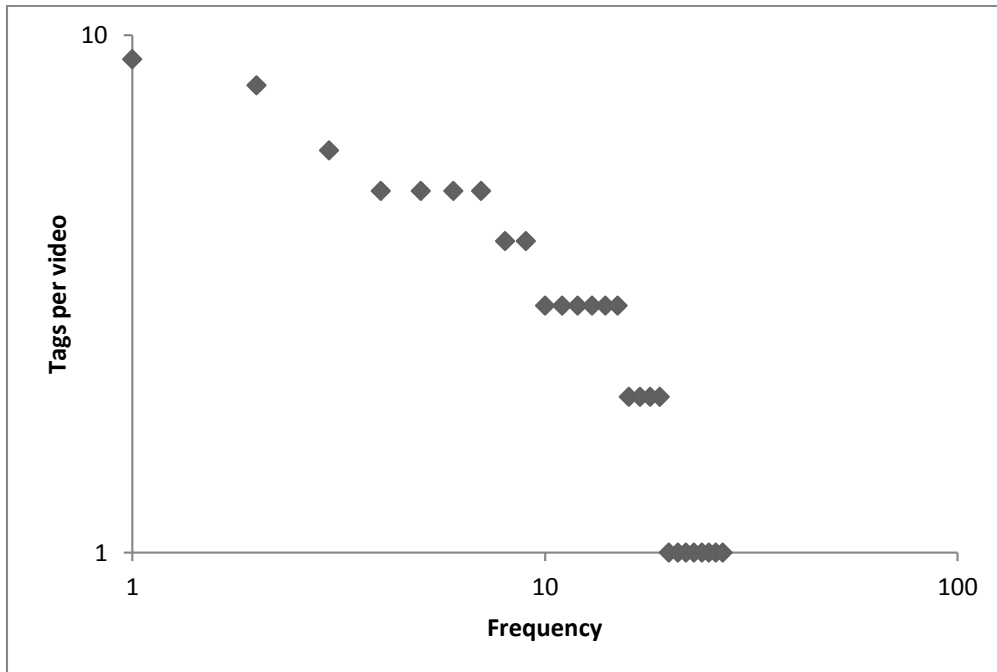


Figure 5-21 Rank frequency of the amount of tags per video.

5.7.3 Discussion

The main promotion was on social media which is increasingly accessed using hand held devices (Lenhart et al., 2010). As the game was not available for hand held devices this will have deterred prospective ‘just try it out’ players. Future versions of video tagging games need to find a method of working on small screen sizes for hand held devices. HTML 5 should also be used due to the lack of support for Flash. This is a challenge as a lot of information needs to be available on the interface without deterring from viewing the video. The change in layout and increasing instructions of how to play did not encourage any more users than in phase one. Some SciFest users still showed little understanding of purpose and needed verbal instructions. Whilst the specific content attracted a few users that had not played phase one, they mostly used the guest account. Whilst interest in content is obviously

important, prospective users need to feel that there is a purpose behind tagging the videos, VideoTag does not offer this in its current format. SciFest users were extrinsically motivated to participate in exhibitions at the one day event. In contrast, attracting users via social media and promotion on the web is difficult due to competition from other websites, games and time fillers. Why would people choose to use VideoTag when they could play Candy Crush Saga or watch a video on YouTube? User opinions of VideoTag in this context will be discussed further in Chapter 6.

Users chose the same few videos to tag despite having 14 to choose from and this trend continued in phase two despite the change to random ordering of videos, from indexing by popularity. The increase in tags per game in phase two indicates that specific interest content encouraged users to tag for longer. SciFest and phase one had less tags per game on average and a lower maximum number of tags per game. An interest in content has more of an effect on a user's motivation to tag than do the game elements. Golder and Huberman (2005) found that special interest groups are likely to tag in a similar way, creating a shared vocabulary and they are also likely to enter tags using similar vocabulary to search terms. This suggests that phase two would create more basic level tags and yet the results indicate that the opposite is true. As discussed in Chapter 7.

5.8 System Comparisons

5.8.1 Games vs. Non-Game

In total, over a 10 month period users played 526 games of Golden Tag, 483 games of Top Tag and used Simply Tag 81 times. Overall the most games (and Simply Tag

sessions) played by one user over the testing period and each phase was 165, and the majority of these were Golden Tag games. Law and Von Ahn (2009) suggest the *Player Retention Curve* as a measure of play frequency; they found few users played the Tagatune GWAP often with the majority of users playing only once. Figure 5-22 shows the player retention curve for VideoTag. It shows that even though VideoTag did not attract large numbers of participants, the users it did attract were more likely to play many games rather than just one. The mean number of games played per user was 42. VideoTag had one main power user who played 128 games. Interestingly, despite playing the most games the power user did not enter the most tags, the next two most frequent users each entered more tags overall.

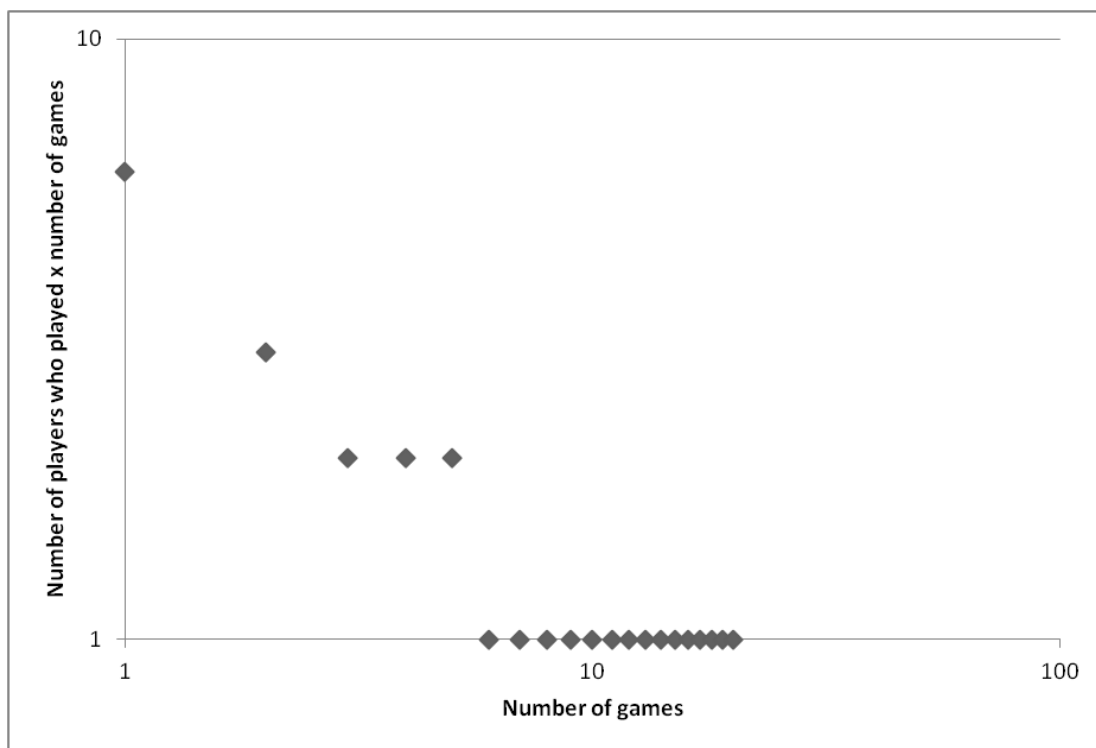


Figure 5-22 Player Retention Curve – rank frequency of the number of games players played.

A total of 2,723 tags were entered into Golden Tag including testing data and each phase. The mean number of tags entered per game was 5, the mode 4, with a maximum of 15. A total of 2,637 tags were entered via Top Tag, the mean number of tags per game was 5, there are two modes of 5 and 6 and the maximum was 18. 929 tags were entered via Simply Tag, the maximum number of tags entered per session was 34, with a mean of 12 and a mode of 1 (see Table 5-13). The majority of tags entered per game in Golden Tag and Top Tag had a frequency of less than 10, 94% of tag frequencies per game were <10 in Golden Tag, 91% in Top Tag. Top Tag shows a slight tendency for users to enter more tags per game. This reflects the findings from SciFest that more users enjoyed playing Top Tag and were more compelled to keep entering tags to find all five top tags. There was more even distribution of tag frequency per session in Simply Tag, with a tendency for users to enter similar amounts of tags per session and at higher average frequency, with users entering less than 10 tags in only 49% of sessions. However, a session in Simply Tag was not restricted to one minute unlike the games. A high maximum number of tags in an individual game would be unlikely due to the time limit. The effect of the time limit is clear from the difference in maximum number of tags per game or session with Simply Tag almost doubling the maximum entered in a game.

Table 5-13 Comparison of the mean, mode and maximum number of tags entered per game in each system.

	Golden Tag	Top Tag	Simply Tag
Total tags entered per system	2723	2637	929
Mean tags entered per game	5	5	34
Mode tags entered per game	4	5 and 6	12
Maximum tags entered per game	15	18	1

Table 5-14 reflects the higher increase in tag agreement of Top Tag users. Most tags relate to the Crazy Science Experiment category used in SciFest, highlighting the impact the success of SciFest had on the datasets. In particular, Top Tag, which was the preferred system for SciFest users, all of the most frequent tags relate to SciFest videos. However, both Golden Tag and Simply Tag only have two tags in the top ten that do not relate to SciFest videos. The majority of tags with high agreement are of a basic level, with the exception of *funny* in Golden Tag which denotes opinion and *advice* in Simply Tag which interprets the content. Chapter 7 will investigate the differences in tag vocabulary further.

Table 5-14 Ten most frequently entered tags in Golden Tag, Top Tag and Simply Tag

Golden Tag		Top Tag		Simply Tag	
Tag	Frequency	Tag	Frequency	Tag	Frequency
man	30	fire	92	water	13
fire	29	water	76	smom	9
glasses	26	bottle	64	fire	7
water	24	man	63	hollister	7
cartoon	19	glasses	39	kids	7
milk	18	liquid	33	advice	6
egg	17	soap	31	bowl	6
funny	15	egg	30	glasses	6
music	14	glass	27	bubbles	5
glass	13	bowl	25	liquid	5

The highest tag frequency in the Golden Tag dataset was 30. Most tags were unique (46.3%) indicating low user agreement. In contrast, for the Top Tag dataset the highest tag frequency was 92 and although the majority of tags were unique (36.9%) there are less unique tags than in Golden Tag indicating higher levels of user agreement in Top Tag. In Simply Tag, 65.3% of tags were unique and the highest tag frequency was only 13. The high numbers of unique tags in Simply Tag could suggest that users entered tags at a more specific level, or it could be indicative of lower use. The difference in tag frequency between the two games cannot be explained by the difference in use as Golden Tag was played more than Top Tag and generated more tags. In Top Tag, where users were encouraged to enter tags of high agreement, users entered the fewest unique tags, suggesting that gameplay affected the tags users entered. Golden Tag gameplay actively discourages agreement, rewarding users for matching low agreement tags and penalising users who enter high agreement tags. A larger proportion of unique tags were entered into Golden Tag but on the whole it seems that the games encourage more higher agreement tags than the non game system. If the Golden Tag gameplay does not encourage users to tag more specifically, then it may create frustration. Figure 5-23 highlights the differences in tag frequency distribution between the three systems, a more even distribution is apparent for Simply Tag in the flatter curve. The steep slopes of Top Tag and Golden Tag are indicative of tagging systems where the majority of tags have low agreement rate. The graph displays the abundance of high agreement tags in the Top Tag dataset compared to the Golden Tag dataset. The predominance of unique tags suggests that users of VideoTag are describers rather than categorisers. The tag classification study will investigate this further in Chapter 7.

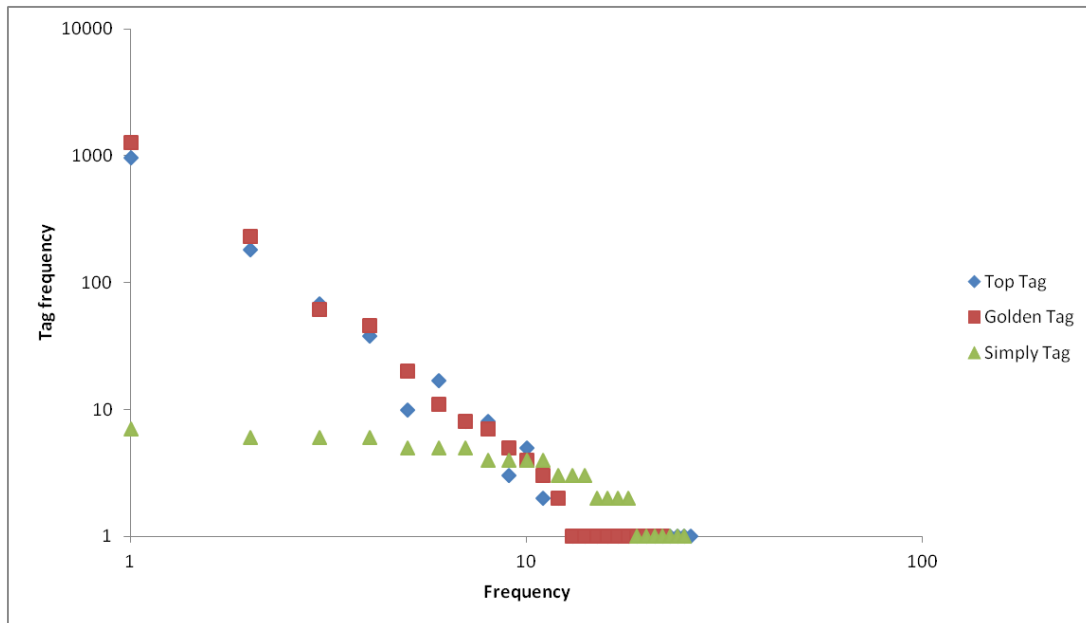


Figure 5-23 Total tag frequency in the three tagging systems.

5.8.2 Phase one Experiment vs. Phase two Experiment

The phase two redesign attempted to give users a greater sense of purpose for using VideoTag and to reduce barriers found in phase one. The redesign included improved communication of purpose and instructions for use on the website, added more specific rather than generic content and provided better integration with an easy fun layer that allowed users to explore the video library and the tag set to find video content. Fewer games and Simply Tag sessions were played and fewer videos were tagged but by more users and more tags were entered. Allowing users the ability to choose which video to tag and providing specific interest videos increased tag frequency and also decreased the number of unique tags. In total, 69.8% of tags were unique in phase two compared to 76% in phase one and 66.9% in SciFest. The amount of tags per game and per video increased in phase two suggesting that an interest in the content encouraged users to enter more tags. The increase in tags per video could be a result of less videos being tagged, but as users were allowed to

choose which video to tag it also highlights that users showed a similar preference for certain videos and categories. Table 5-15 highlights the differences in tag frequency between the three experiments, showing the impact of SciFest on the phase two tag frequency. Low participation affected the tag frequencies of phase one and phase two and comparing the frequencies of phase two and SciFest shows that few tags entered into SciFest were entered by other users during phase two. In both phase one and phase two, Simply Tag generated the highest percentage of unique tags, suggesting that games encouraged higher levels of tag agreement. Gameplay had an effect on the amount of unique tags entered in phase one and SciFest, but not in phase two, where results suggest content had more effect on tag type than game elements. General inspection of the tags suggest that phase two encouraged users to enter tags of higher specificity, further investigation through tag classification (see Chapter 7) will reveal if this is the case.



Table 5-15 The number of high frequency tags entered in each experiment.

Phase 1		Phase 2		Phase2 no SciFest		SciFest	
music	10	fire	120	glasses	11	fire	117
car	8	water	108	smom	9	water	103
game	7	man	82	kids	8	bottle	78
dance	6	bottle	79	tents	8	man	76
men	6	glasses	59	driving	7	glasses	47
song	6	egg	47	guitar	7	liquid	47
funny	5	liquid	47	live	7	egg	46
girl	5	bowl	42	advice	6	bowl	42
rap	5	glass	42	dwight	6	glass	39
animal	4	soap	41	festival	6	soap	39

5.8.3 Motivation

The results indicate a lack of motivation to use VideoTag. Table 5-16 details situational motivation states, describing how users could be attracted to VideoTag in each state. The model was developed from the concepts of motivational theory of Ryan and Deci (2000), Vallerand (1997) and Kowal and Fortier (1999). Likelihood of flow is dependent on motivation type. If high instances of intrinsic or self-determined extrinsic motivation are found then there is higher probability of users engaging in the game/system. More instances of non-self-determined extrinsic motivation suggest a lack of engagement. The majority of users were non-self-determined and influenced by extrinsic motivation. These users are the least likely to enter a state of flow and immerse themselves in the system. VideoTag was unsuccessful at attracting intrinsically motivated users. The system did not offer users high levels of perceived use, and game elements were not enough to engage users on their own as they were too tightly integrated to the tagging process. Users with Extrinsic Identification and Extrinsic Introjection motivation were the most prolific users in the VideoTag system. User perception of purpose and their level of enjoyment were not enough to integrate extrinsic motivation and transform users into intrinsically motivated taggers.

Table 5-16 User motivation to use VideoTag

Situational Motivation	Self Determined	Intrinsic	User Types	Likelihood of Flow
			<p>Perceived usefulness for the tags.</p> <p>Interest in tagging videos.</p> <p>Interest in video content and perceived purpose in uploading content.</p> <p>Desire to create a community.</p> <p>Desire to conquer the site, either to break it or score as many points as possible, potentially by cheating.</p>	<p>Most Likely</p> 
Non Self Determined	Extrinsic	Integration	<p>If elements of the games attract users in identification state they will play more.</p> <p>If users find content they enjoy watching.</p>	
		Identification	<p>Interest in gamification</p> <p>Interest in GWAP</p> <p>Interest in specific interest video content</p> <p>Researching similar</p>	
	Introjection	<p>A feeling of obligation towards helping the author.</p>		
	External Regulation	<p>Request to participate in user study;</p> <p>Financial reward for registering.</p> <p>Visitor to SciFest</p> <p>Points and Levels</p>	<p>Least Likely</p> 	

5.9 Conclusions

The majority of traffic for the first VideoTag project (Greenaway, 2007) came from a blog post by Mashable.com that was reproduced by other tech bloggers. The Greenaway (2007) project was released at the birth of web 2.0 when many bloggers and technology press were excited about any new Web 2.0 project. Six years on Web 2.0 is now a part of daily life. The landscape of the internet has changed and it is no longer desktop based web apps getting publicity but instead apps for smart phones and tablet computers. The app market is saturated with games and other services and so it is extremely difficult for a project to gain visibility to a mass audience. A system like VideoTag needs to have all the factors necessary for technology adoption, perceived usefulness and perceived ease of use. To invite play there must also be perceived enjoyment, a set of clear goals and adequate challenge for players to accomplish. It was difficult to create interest in the project and secure participants. Video tagging games do not currently have mass appeal as a game or form of entertainment but they can provide a layer of entertainment in a system that is perceived as useful.

The finding that users preferred to tag in a game environment shows that users can be encouraged to tag videos if games are provided. However, to be enticed to use the system they must feel a purpose in tagging the videos. The low level of participation suggests that VideoTag did not offer prospective users enough perceived usefulness. There was no personal benefit for the tags they entered as a result users lacked intrinsic motivation to engage in VideoTag. The game elements and reward structure provide extrinsic motivations for use, but these proved insufficient to motivate enough users. Without intrinsic motivation no amount of extrinsic motivation will encourage sustained use. However, the users who did use VideoTag were more

likely to play many games than play one game and leave, which shows if users are motivated to use the system, either game elements or interest in content encouraged them to tag for longer. Results of phase two suggested that if users are attracted to VideoTag for specific content, they preferred to watch the whole video and tag in a non-game environment. Further research in phase two investigated whether users with a specific interest in the video content are more motivated to tag in a non-game environment. The findings suggest that interest in content is more of a motivator over game elements. The SciFest results suggested without an interest in the content game elements encourage participation in a more controlled environment, but were not sufficient to sustain use. Users in SciFest were extrinsically motivated to use VideoTag because it was an exhibition in the event they were attending. Unfortunately, users remained externally regulated; VideoTag design features were insufficiently motivating to create regular players.

The phase two redesign addressed numerous issues highlighted by the phase one experiment and also the phase two prototype at SciFest. Improving sense of purpose, adding step-by-step instructions to get started with VideoTag and adding access to the video library to improve perceptions of use did not sufficiently increase the numbers of users. As the number of unique visitors was similar to the number in phase one, the phase two promotion apparently failed to reach enough potential players. The methods applied to improve perceived usefulness were insufficient, as users who were attracted were not adequately motivated to interact with the system. Adding the guest account was beneficial for only one select group of users with a specific interest. This indicates that using specific interest content is a good way to motivate users to use VideoTag but this particular group of users were not interested in the games or the purpose of tagging. Putting emphasis on video content attracted

participants that were motivated to just watch the videos and they had little interest in tagging. A further deterrent was that the system was optimised for desktop browsers. Whilst the system was designed to be viewed on a smaller screen, unfortunately many devices ended support for Flash player, which is used to embed the YouTube videos. Twitter was used as the main vehicle for promotion but a large proportion of Twitter users access the service using hand held devices like smart phones or tablets that could not access the content. HTML 5 can be used to embed the videos, but this is only part of the solution; future implementations of VideoTag need to consider how the interface can be transformed to be functional on hand held devices. User studies discussed in the next chapter highlight more areas for improvement and the aspects of the system that participating users enjoyed.

6 Usability and Engagement Evaluation

6.1 Introduction

Usability testing for software and web development focuses on identifying problems in the system that can impede the user's ability to complete actions. Usability testing is conducted in two ways: heuristic evaluation and user testing. Heuristic evaluation allows the development team to look at the interface and pass judgement on its usability based on a set of guidelines. User testing involves observing or monitoring participants as they use the system and the completion of questionnaires by system users. Both methods identify problems that inhibit a user's ability to complete actions within the system. Generally, heuristic evaluation can identify and correct major issues before users test the system (Nielsen and Molich, 1990). Heuristic evaluation offers a quick and inexpensive method of usability evaluation compared to employing participants for user tests. It can be used throughout the design stage, starting before a prototype has been developed. For optimum performance, however, heuristic evaluation should be used in conjunction with user tests (Schaffer, 2008).

Usability, as defined by ISO (1998), should cover the effectiveness and efficiency of the system and user satisfaction (Brooke, 1996; Barnum and Palmer, 2010) and incorporate memorability and reliability (Orehovacki, 2010). Key principles for designing good usability that have been adopted as best practise have been proposed by two authors, Shneiderman (1987) and Nielsen (1994). Shneiderman (1987) produced eight golden rules for interface design:

1. Strive for consistency - use identical terminology, consistent colour and layout;

2. Cater for universal usability – cater for all ages and abilities;
3. Offer informative feedback – provide users with feedback for actions so they know if it has been successfully executed;
4. Design dialogs to yield closure – all actions need a beginning, middle and end, informing users when the end is reached;
5. Prevent errors – reduce the amount of errors a user can make and allow users to recover from errors;
6. Permit easy reversal of actions – allow users to easily undo or go back after every action;
7. Support internal locus of control – allow users to feel in control of the system rather than that their actions are controlled by it;
8. Reduce short term memory load – do not force users to remember information from one screen to the next.

Nielsen and Molich (1990) developed these rules to create ten usability heuristics for interface design that system designers and developers should follow to create usable systems. Many usability problems were elicited then through factor analysis, creating nine categories of usability problem. After ongoing refinement of these categories, Nielsen (1994) presented the following ten usability heuristics :

1. Visibility of system status – keep users informed of what is going on;
2. Match between system and real world - use natural language all users can understand;

3. User control and freedom – allow users control over the system and the ability to reverse actions;
4. Consistency of standards – use identical language, colour and layout so users can distinguish between same and different actions;
5. Error prevention – prevent errors from occurring as much as possible. Predict errors and present users with a confirmation option;
6. Recognition rather than recall – do not make users remember information from one screen to the next, let them identify what to do, not remember instructions;
7. Flexibility and efficiency of use – allow expert users the ability to speed up actions, create shortcuts;
8. Aesthetic and minimalist design – only display relevant information that a user needs to complete actions;
9. Help users recognise, diagnose and recover from errors – use plain language error messages that users can understand and suggest solutions;
10. Help and documentation – a system should be intuitive without the need for documentation, but help should be available if a user needs it.

In terms of web development, these guidelines were created for an internet landscape very different from the one in use today and, if followed stringently, many successful websites (e.g., YouTube and Facebook) would seem to have bad usability (Silva and Dix, 2007). There has been a move to design for the user experience with factors that focus on playful interaction, social interaction and overall enjoyment

(Shneiderman, 2004). An early pioneer for designing for enjoyment, Malone (1982), proposed a set of heuristics for designing enjoyable user interfaces. His categories of heuristics were: *Challenge*, *Fantasy* and *Curiosity*. *Challenge* heuristics consider the goal of the system. Goals must be clear; there may be multiple goals at varying levels of difficulty. The user must be presented with regular feedback indicating how close they are to achieving their goal(s). *Fantasy* heuristics refer to the fiction (Juul, 2009) and the aesthetics; they should be emotionally appealing and embody metaphors that model the real world and that are familiar to the user. *Curiosity* heuristics cover attributes of design that invigorate a user and promote continued use. The heuristics recommend the use of audio or visual effects to enhance the fantasy or to provide feedback ("juiciness" - (Juul, 2009)), to keep the users well-informed of what the system offers and providing new information if a user's knowledge is incomplete. The interface should also use randomness to add variety and humour.

User experience and enjoyment can be measured using the theory of flow developed by Csikszentmihalyi (1975), as discussed in detail in Section 2.3.3. The usability heuristics described previously share some similarities with the characteristics of flow: clear goals, feedback on actions and control over actions, which includes the ability to recover from errors. Flow also describes a confidence using the system: the user's perception of their own skills matching the challenges that the system sets. All conditions of flow must exist to enable the flow state, defined as complete immersion in an activity so that all sense of time and state of being is lost. Conditions of flow provide the basis for game design theory, and so to create flow state for a player is the ultimate goal for any game designer. Computer games share some of the same barriers to use as other computing systems. One difference is that usability that is too good can take something away from a game. Good usability for games has different

goals than good usability for software or a website (Lazzaro, 2008). Whilst a game should be easy to learn but difficult to master, software or a website “should be both easy to learn and easy to master” (Malone, 1982, p.66). Usability problems form part of the challenge for players of hardcore games whereas they will deter casual game players (Fortugno, 2008).

There are no set guidelines for game usability or playability evaluation as there are for usability in software or web development (Laitinen, 2008). Laitinen (2006) found large numbers of usability problems in games using Nielsen's (1994) heuristics, but they did not highlight problems with player enjoyment. Schaffer (2008) claims that tailoring usability heuristics for games will help to find experience-orientated problems rather than task-orientated problems. The first set of usability heuristics for games was produced by Federoff (2002). Schaffer (2008) claims that although they created a basis for further research, many heuristics were too general and too heavily focused on game design theory and not on identifying usability problems. Desurvire *et al.* (2004) propose HEP – ‘Heuristic Evaluation for Playability’. These heuristics are split into four categories: *Gameplay*, *Game Story*, *Game Mechanics* and *Usability*. Sixteen heuristics in *Gameplay* refer to how the game is played, user interaction with game elements, pace, rewards, control over actions, match between challenge and skills and difficulty. *Game Story* heuristics relate to the fantasy (Malone, 1982) of the game. Eight heuristics measure how clear and attention-grabbing the narrative is, how players relate to game characters and whether they feel that they have control over the outcome of the game. Seven *Game Mechanics* heuristics refer to how the game is built and how users interact with the interface, adapting standard usability guidelines for game design, such as clear goals, easy to learn and intuitive controls. Twelve *Usability* heuristics adapt and extend Nielsen's (1994) usability heuristics. The

performance of the HEP heuristics was tested by evaluating a prototype game and comparing the results to playability user tests by observing users playing the prototype and asking them to complete a satisfaction questionnaire. Desurvire *et al.* (2004) found that the HEP highlighted more usability problems but that the playability user tests gave more specific information about the issues from a user perspective. The authors claim that for optimum performance the heuristics need to be combined with user studies because heuristic evaluation can never predict user behaviour. HEP is best suited for evaluating general issues in early development with a prototype. Further limitations of HEP are that it is only useful in limited circumstances and needs modifying for each individual game.

Barendregt *et al.* (2006) combine Nielsen's (1994) usability heuristics with Malone's (1982) enjoyable interface heuristics to highlight problems for use, interaction and fun in computer games. Usability problems with games are split into four categories: *Knowledge* - there are no clear goals; *Thought* - bad navigation means that users can not complete actions; *Memory* - high cognitive load creates interactions that rely on a user's memory; and *Judgement* - feedback is unclear and open to misinterpretation. Interaction problems are defined as *Habit* - the correct action is performed in the wrong situation; *Omission* - a user does not complete all areas and *Recognition* - feedback is not noticed or confused. Fun problems are categorised as *Challenge Fantasy*, *Curiosity* and *Control*. They highlight problems that make the game less motivating to use, such as the challenge level being too high or too low; a user may not like the graphics or the narrative (e.g., too violent or too childish); a user can become frustrated if there is not adequate level progression and become impatient if they feel that the game has control over their actions. Barendregt *et al.* (2006) found that most problems revealed by the heuristic evaluation were connected to a lack of

knowledge about how to play the game. They re-evaluated users after they had practised the game and found that many problems were different to the problems that occurred during their first use. They suggest that user testing should be extended to include re-tests over time.

Whilst usability and playability heuristics are useful for understanding what aspects of a game might motivate or demotivate users, there are no methods to evaluate user enjoyment, which is a key to the success of any game. Enjoyment is a result of engagement; users need to engage with elements of the game in order to enjoy the playful experience. Engagement has been measured in other contexts. For example, O'Brien and Toms (2010) developed the User Engagement Scale, a set of heuristics to measure user engagement in the information retrieval process. These heuristics are based on Nielsen's (1994) usability heuristics: perceived usefulness, ease of use and conditions of flow. To evaluate engagement is not to evaluate the presence of flow; flow is an optimum condition and an enjoyable experience can be had without the presence of flow (Turner, 2010). Engagement is created via an interest in the content of the game and by the game's ability to hold the player's attention (Chen et al., 2011), as well as by the perceived usefulness of the game (Turner, 2010). Users will engage with a positive emotional experience and disengage if it is negative (Turner, 2010). Any enjoyment or engagement evaluation therefore has to predict whether the gameplay, game story, game mechanics and game usability will produce positive or negative emotions. As emotion is subjective, this makes enjoyment and engagement difficult to model. Bartle (2009) claims that it is difficult to capture why people play when you can only see what they are doing. To be engaged with a GWAP a user must feel competent at completing the tasks, feel related to the project in some way and be determined to perform well (Ryan and Deci, 2000). For engagement to occur the mechanics, dynamics and aesthetics (Hunicke et al., 2004) must effectively support the player. Sweetser and Wyeth (2005) argue that player enjoyment is the

single most important goal for a game. If players do not enjoy a game they will not play. People usually play games for the game itself, without external rewards. People who get the most out of the game playing experience are thus intrinsically motivated to play the game. If external reward is required to encourage people to play a game, for example offering a financial reward for playing a GWAP, then they are unlikely to immerse themselves fully into the game.

Sweetser and Wyeth (2005) propose a model, called Game Flow, which can be used to evaluate player enjoyment. It models games based on the elements that are conducive to a flow state, flow being the recognised measure of enjoyment in games, rather than heuristic evaluations. The model contains eight categories: *Concentration, Challenge, Skills, Control, Clear Goals, Feedback, Immersion* and *Social Interaction*, each contain individual heuristics that can be used to evaluate player enjoyment. The model, whilst applicable to all games, has a bias toward MMORGS (Massively Multiplayer Role Playing Games). Despite this, the categories derived from flow theory are relevant to all games and whilst some of the criteria are specific to MMORGS, these can be marked as not applicable without affecting the overall evaluation. There is also scope for other more suitable criteria to be added. Sweetser and Wyeth (2005) claim that the model is not meant to be an evaluation tool for game developers in its current format but that it is a useful tool for reviewing games and indentifying issues as well as the affect that these issues have on player enjoyment. They found it easier to identify what went wrong than what was done well.

Febretti and Garzotto (2009) compared user studies using customised usability heuristics based on Nielsen (1994) and custom playability heuristics based on

Sweetser and Wyeth (2005). The results were analysed to find correlations between usability and engagement and playability and engagement. They discovered a statistically significant low correlation between usability and long term engagement in a game and that playability heuristics had a significantly higher correlation with engagement than with usability. In particular, playability heuristics relating to *Challenge* had the highest correlation with engagement. The highest correlation for usability was with heuristics relating to *Control*. The findings corroborate the theory that flow state is a condition of engagement; to achieve flow a user must be in control of their actions and their perceived ability must match their perception of the challenge. The second highest correlating playability heuristic was social interaction. Whilst this does not appear in the eight flow conditions it is essential in promoting user engagement in games (Sweetser and Wyeth, 2005).

This chapter discusses three user studies of the VideoTag system. For each study the evaluation methods are described. Results are then analysed, focussing on the reliability of the measures as well as the significance of the results. Finally, an expert evaluation is conducted using the Sweetser and Wyeth (2005) Game Flow model to measure enjoyment. A discussion of the findings with comparisons with user studies research concludes the chapter.

6.2 Usability Evaluation

6.2.1 Methods

The effectiveness of a video tagging game is measured by the user's ability to complete tasks and the quality of their output (Brooke, 1996). Quality of output will be assessed through tag classification studies (see Chapter 7). The efficiency of the

system, rather than the amount of games played, is better measured by *Throughput* and *ALP*, as defined by Von Ahn and Dabbish (2008) is discussed in Section 2.3 on p29. User satisfaction is measured through a usability questionnaire. Tullis and Stetson (2004) compared five different usability questionnaires: their own questionnaire and four established methods (QUIS, SUS, CSUQ and Microsoft Product Reaction Cards). The usability of two websites was evaluated and the responses to the five questionnaires compared. Between 19 and 28 participants completed each questionnaire. To calculate accuracy for the smaller numbers of participants that are more common for user tests, sub-samples of 6 to 14 questionnaires were also taken. The sub-sample data was compared to the original response data using a t-test. They discovered that the accuracy of all questionnaires improved with larger sample sizes, recommending a sample size of 12-14 participants. The SUS scale proved to be the most reliable and to have the highest degree of accuracy, particularly with small sample sizes. It was the only questionnaire that achieved 100% accuracy, found with sample sizes over 12.

Based on the findings of Tullis and Stetson (2004) the Simple Usability Scale (SUS) was chosen to measure the usability of the VideoTag system. SUS is a simple ten item scale measuring user satisfaction using a Likert five-point scale, where 5 is strongly agree and 1 is strongly disagree. The SUS scale was created by Brooke (1996): Fifty potential questions were tested with one group of users evaluating two systems. After a correlation analysis the ten questions with very high positive and negative agreement were retained. Positive and negative questions alternate in the final scale, with odd numbered questions being positive and even numbered questions being negative. This is to avoid acquiescence bias and artificially high values (Sauro and Lewis, 2011; Barnum and Palmer, 2010). Sauro and Lewis (2011) warn of reliability

problems due to users accidentally agreeing with negative items and miscoding by researchers. They investigated the benefits of using an all positive version of the SUS scale but found no significant improvement in user scores or any evidence of acquiescence bias in the original SUS scale.

The VideoTag usability evaluation was uncontrolled; the questionnaire was hosted on the VideoTag website⁶ but was not linked to internally. A link was emailed to members of the Statistical Cybermetric Research Group within University of Wolverhampton with simple instructions:

Step 1 – Sign up for a VideoTag account;

Step 2 – Spend at least 10 minutes playing Golden Tag;

Step 3 – Spend at least 10 minutes playing Top Tag;

Step 4 – Spend at least 10 minutes tagging in Simply Tag.

The link to the questionnaire was also sent to VideoTag users who had indicated a willingness to participate and posted on Facebook and Twitter using both VideoTag specific (@videotag2 on Twitter and on the VideoTag Facebook page) and shared on personal accounts of the author. This method was used for the all three questionnaires discussed in this chapter. See Appendix C for the questionnaire.

⁶ <http://www.videotag.co.uk/sus-ue.php>

6.2.2 Results

Five participants completed the SUS questionnaire. Initial qualitative analyses of the data revealed that the majority of participants agreed that VideoTag was easy to use, they could accomplish tasks without help, it was consistent, functional and fit for purpose; however, most agreed they were unlikely to use VideoTag frequently. A total SUS score was calculated using a method described by Brooke (1996) which converts each individual score of 1-5 into an individual score with a range of 0-100. This creates what the author describes as “a composite measure of the overall usability of the system” (Brooke, 1996, p.5). As scores for individual items are meaningless (Brooke, 1996), no analysis of the frequency of individual scores or individual means was conducted. Before evaluating the SUS scores, especially considering the small sample size, a statistical test for reliability of the scale was conducted.

H0 - SUS scale is not a reliable measure of usability for VideoTag.

A factor analysis could not be conducted because there were only 5 participants and Nunnally (1978) recommends at least 10 for a factor analysis. Usually factor analysis is conducted in conjunction with Cronbach’s alpha to assess the reliability of the scale across interrelated groups of questions in questionnaires. In this instance only Cronbach’s alpha could be used. In order to calculate Cronbach’s alpha scores correctly and avoid miscoding (Sauro, 2011) responses for negatively phrased questions were converted to the positive scale. This avoids an unrealistic negative alpha score (Field, 2013; Sauro, 2011; Yu, 2001). Cronbach’s alpha was $\alpha=.92$, which is high. Results over $\alpha=.7$ are considered reliable (Nunnally, 1978). Redundancy was found with questions 5 and 10. This could indicate high interrelatedness between

variables (Cortina, 1993) or could be a result of a high number of items and low number of participants.

Inter-item correlation revealed a negative correlation between questions 6 and 7. If a user thinks there is high inconsistency on the website then they are most likely to think there is less chance of other users learning to use the website quickly. No correlation was found between questions 5 and 6. The relationship between responses to individual questions was investigated further with a Kendall's Tau correlation. Kendall's Tau is a non-parametric test that measures the association between two quantities and it is recommended for small sample sizes over a Spearman or Pearson correlation (Field, 2013). No correlation was found between questions 5 and 6 or 6 and 7 and a negative correlation was found between questions 10 and 5, indicating that removing questions 10 and 5 would improve reliability of the survey. The lack of a linear relationship between 5 and 6 and 6 and 7 had an effect on the Cronbach's alpha calculation. Question 9 has a significant relationship with questions 1, 3 and 8: feeling confident using the website is related to whether a user will use the website frequently and ease of use. The impact of this positive correlation can be seen in Table 6-1 and so removing question 9 would reduce the reliability and removing the questions with no correlation would reduce redundancy. The Cronbach's alpha score of $\alpha=.92$ gives evidence that SUS is a reliable scale for measuring VideoTag's usability.

Overall usability must be analysed by calculating the sum of scores for each item (Brooke, 1996). Each item's score was converted to a 1 to 4 scale. Odd items score the scale position minus 1 and even items score 5 minus the scale position. The sum of

scores was multiplied by 2.5 to convert it to a 0 to 100 scale (Brooke, 1996). The original description of the SUS scale by Brooke (1996) does not provide a method to interpret the SUS score. Subsequent research has suggested that any score above 70 indicates good usability, and that below 70 suggests usability problems. Bangor *et al.* (2009) suggest that >90 is exceptional, >80 is good and >70 is acceptable. The mean score is 71, and the median is 80, signifying that users thought VideoTag had good overall usability. Figure 6-1 shows the range of SUS scores. Whilst the lowest extreme could be an outlier, this cannot be assumed due to the small sample size.

A qualitative inspection of individual scores ruled out user mistakes. There was no evidence that the negative user misunderstood the scale as a range of responses for both positive and negative questions were entered. The overly positive scores do not have enough variety to rule out positive bias, however. Recalculating the mean without the overly positive user's score reduces the mean to 64.34 and the median to 68.75. Whilst this mean and median may be more accurate, the overly positive score cannot be ignored as there is no definite evidence that they have not entered genuine responses. The overly positive user's score does not reduce the mean enough below 71 to indicate that usability problems are being missed.

Table 6-1 Results of the Cronbach's alpha test for reliability for the SUS questionnaire.

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q1-p	35.20	83.700	.903	.907
Q2-n	33.80	84.200	.875	.909
Q3-p	34.60	79.300	.917	.904
Q4-n	33.40	88.800	.795	.915
Q5-p	34.00	95.500	.333	.932
Q6-n	33.80	93.200	.407	.929
Q7-p	33.60	91.300	.638	.920
Q8-n	34.40	69.300	.921	.904
Q9-p	34.60	74.800	.987	.898
Q10-n	34.60	81.300	.562	.930

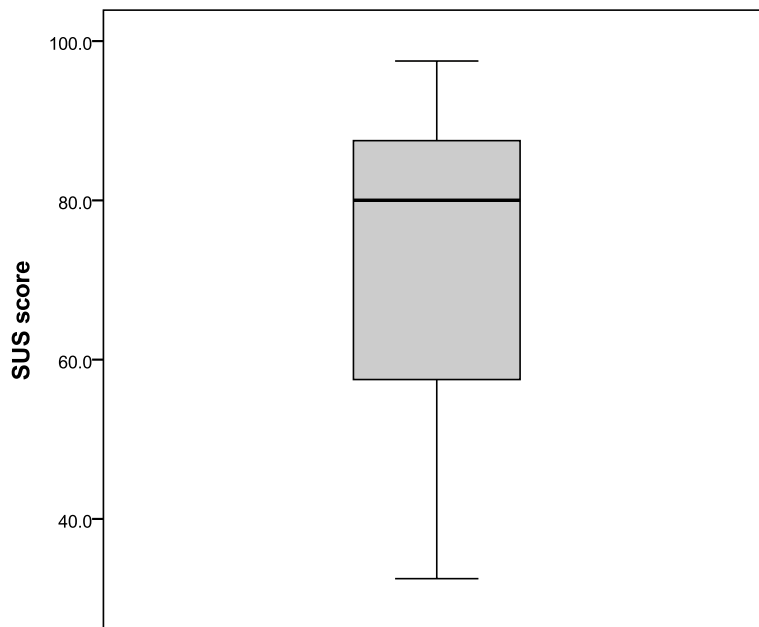


Figure 6-1 A box plot of SUS scores.

6.3 User Satisfaction

6.3.1 Methods

The SUS scale is effective at measuring user opinion about general usability issues to highlight any usability problems, but it does not have the scope to gauge user opinion about their experience of using the VideoTag system. Playability (see 6.4) and enjoyment (see 6.5) evaluations assess user interaction with game elements; this study uses Microsoft Product Reaction Cards (MPRC) to gauge user opinion of their interactions and reactions with the whole system (Johnson, 2012). Tullis and Stetson (2004) included MPRC in their comparison of usability evaluation methods; although the method is not as accurate or reliable as the SUS scale, Barnum and Palmer (2010) recommend triangulating MPRC with other questionnaires to complement the findings. MPRC as a method to measure desirability was developed by Benedek and Miner (2002) and extended by Williams *et al.* (2004). Through market research, user

research and team brainstorming Microsoft researchers selected 118 words. Based on evidence that users are more likely to agree on positive descriptions than negative, a 60/40 positive to negative ratio was chosen (Barnum and Palmer, 2010). MPRC offers a user-centred method for capturing user satisfaction at low cognitive cost; the user selects words to enter rather than being asked their opinion. This allows for control over the subjective vocabulary used and enables frequency analysis. This method was chosen to analyse VideoTag because of its obvious correlation with the tagging process. Participants were recruited following the methods described in 6.2.1; the words were presented to the user on a webpage⁷, see Appendix C for the full list of words.

6.3.2 Results

A total of 73 words were entered by six participants and 47 of these words were unique, forming 40% of the 118 words presented to users. Users recorded opinions relating to ease of use and game experience. Words with high agreement representing ease of use include *Accessible, Usable, Clear, Intuitive, Responsive, Straight-Forward* and *Understandable*. Words with high agreement depicting game experience include *Fun, Novel* and *Entertaining*. Table 6-2 shows the selected words and the frequency at which they were entered.

The low number of participants limits the amount of analysis that can be conducted. Fewer participants gives a lower probability of users agreeing on words. Without a

⁷ <http://www.videotag.co.uk/mprc-ue.php>

large range of user opinions being harvested the responses are more subjective and conclusions based on user agreement are more difficult to form. Figure 6-2 plots the frequency distribution of the words. More low rank, high agreement words would be expected in a larger sample population as well as many more high rank, low agreement words. For user experience evaluation the words that users agree on hold the most information for the analysis. The low ranking, high agreement words (frequency ≥ 3) recorded in this sample relate to the usability of the system (e.g. *Easy to use*, *Accessible*, *Usable* and *Understandable*) and only one low rank, high agreement word relates to enjoyment of the system, *Fun*. All negative words had high rank, low agreement with the exception of *Undesirable* and *Overwhelming*, which were selected by two users.

Table 6-2 Frequency of words entered in the MPRC evaluation.

MPRC word	Frequency
Easy to use	4
Accessible	4
Fun	3
Usable	3
Understandable	3
Approachable	2
Attractive	2
Clear	2
Entertaining	2
Friendly	2
Intuitive	2
Inviting	2
Novel	2
Overwhelming	2
Responsive	2
Stable	2
Stimulating	2
Straight Forward	2
Undesirable	2
Appealing	1

Boring	1
Complex	1
Confusing	1
Consistent	1
Creative	1
Customizable	1
Difficult	1
Engaging	1
Enthusiastic	1
Essential	1
Gets in the way	1
Hard to Use	1
Impressive	1
Innovative	1
Predictable	1
Relevant	1
Simplistic	1
Stressful	1
Time-consuming	1
Trustworthy	1
Unattractive	1
Unrefined	1
Valuable	1
Annoying	1
Comprehensive	1
Creative	1
Meaningful	1

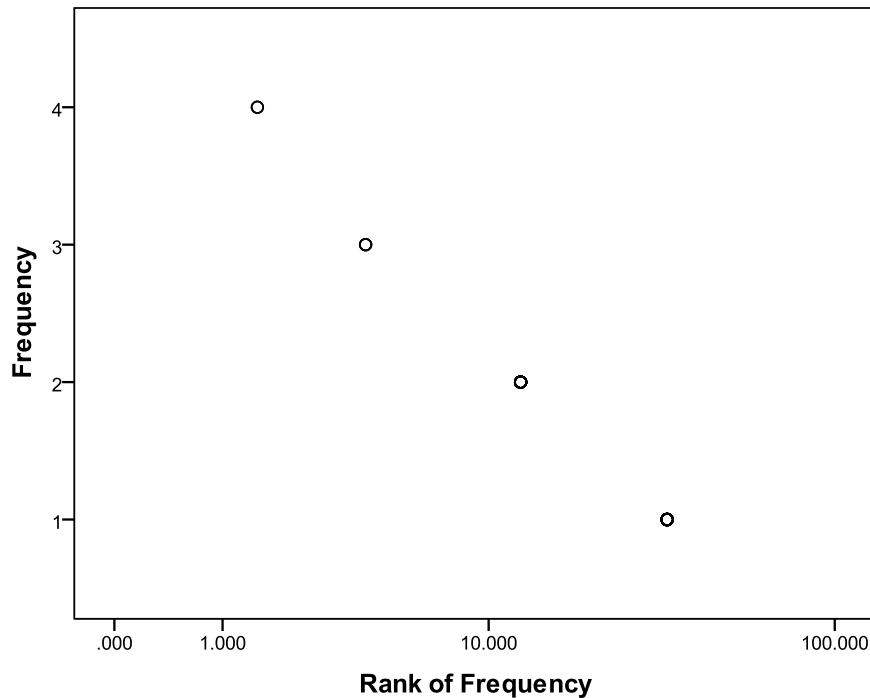


Figure 6-2 Frequency Distribution of MPRC words

To facilitate further analysis, words were grouped into two categories: Positive and Negative (pos/neg) (Johnson, 2012; Tullis and Stetson, 2004) and Enjoyment and Usability (E/U). Johnson (2012) used four categories, initially: Appearance, Judgement, Emotive and Use. However, after a factor analysis revealed that there were really only two categories, these were amended to Quality and Use. Word groupings are subjective and relative to the system being evaluated therefore only loose groupings of words can be created.

Figure 6-3 shows the distribution of positive and negative words; 33 positive and only 14 negative, 70.2% and 29.8% respectively. There are no low rank, high agreement negative words. This supports Barnum and Palmer's (2010) findings that MPRC tests produce more agreement with positive terms. Of the 14 negative words, 7 were entered by one user, 50% of the total; of the 33 positive words, 17 were

entered by one user, 51.5% of the total. Although the user studies were anonymous, on comparing the timestamps of users and SUS results, the results suggest that two users had extreme opposing opinions of VideoTag. With such small participant numbers these extreme opinions, accurate or not, skew the results. With a larger population sample it would be possible to remove them as outliers.

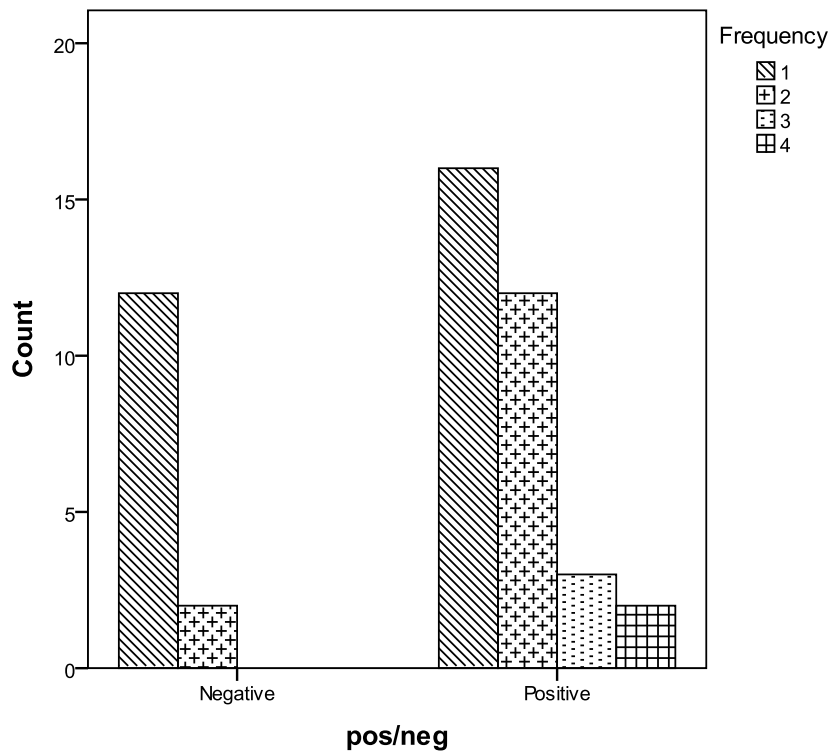


Figure 6-3 Frequency of positive and negative MPRC words.

Figure 6-4 depicts the frequency of words related to enjoyment and usability: 29 enjoyment and 18 usability, 61.7% and 38.3% respectively. Overall there are more words entered relating to enjoyment than to usability, but more agreement on words relating to usability than enjoyment is evident; usability also has a more even distribution. Usability has a more standardised vocabulary (i.e. *usable*, *ease of use*,

accessible, intuitive) whereas enjoyment is more subjective, with less users agreeing on descriptions of their experience. Enjoyment words make up the majority of the long tail or high rank, low agreement words.

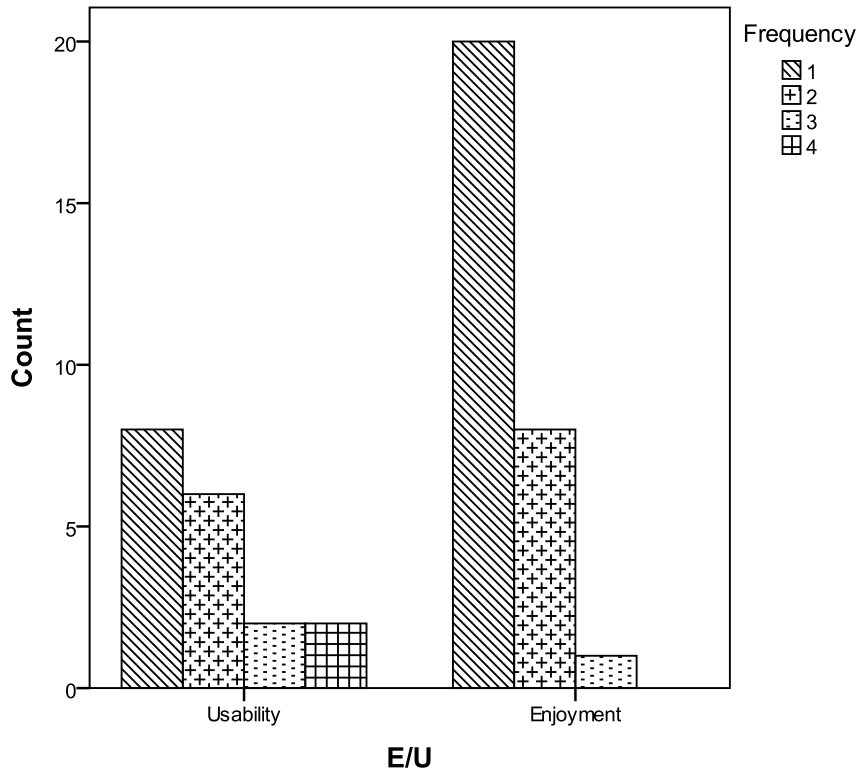


Figure 6-4 Frequency of enjoyment and usability MPRC words.

Figure 6-5 compares the distribution of positive and negative words in the usability and enjoyment categories. Only 3 words were entered for negative usability compared to 18 words for positive usability. There is a smaller difference in the enjoyment category, with 11 words entered for negative enjoyment and 18 for positive enjoyment. Both extreme users entered more words related to enjoyment than to usability.

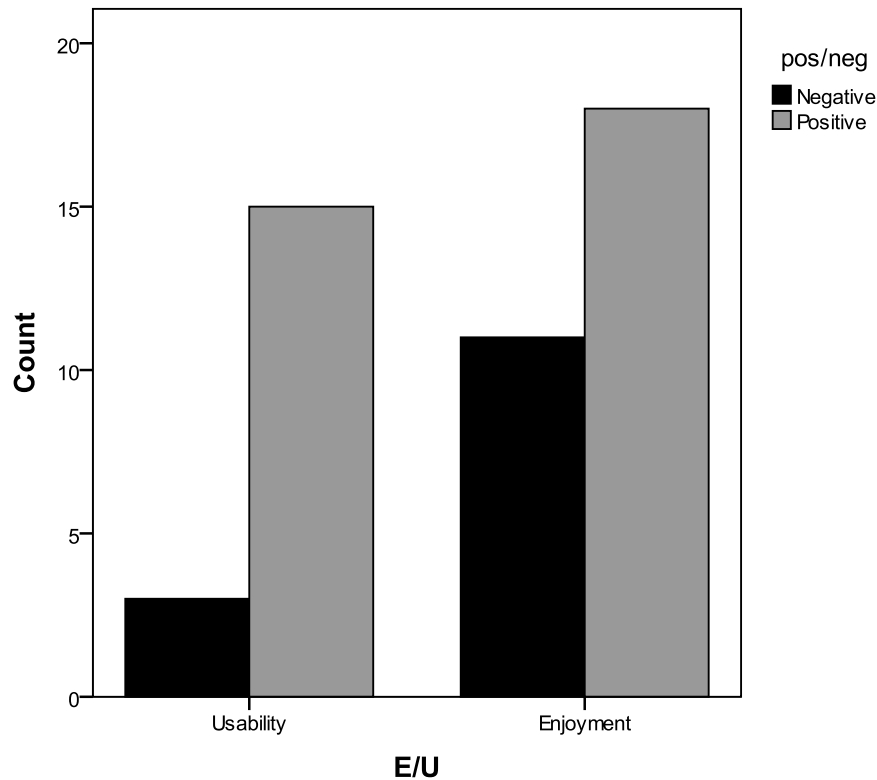


Figure 6-5 Amount of positive and negative MPRC words grouped by enjoyment and usability.

To evaluate whether there was a difference in the distribution of words in the positive and negative and the enjoyment and usability categories, two chi square tests were conducted. The frequencies of words in each category were compared to the frequencies of words in the overall sample. No significant difference was found between the distributions of positive and negative words ($p=.060$). A significant difference was found between the distribution of words in the enjoyment and usability categories ($p=.000$).

Overall there was more agreement that VideoTag had good usability than that VideoTag was enjoyable. A high proportion of words were entered that suggest that

whilst VideoTag is easy to use, it has problems that have a negative effect on enjoyment. The words alone only describe a reaction and they are not related to a specific feature of the website or games.

6.4 Playability

6.4.1 Methods

Playability evaluation was conducted to measure user opinions about how effectively game elements have been integrated into the video tagging system. A playability questionnaire created by Goh *et al.* (2011) was chosen for its relatedness to GWAP. Goh *et al.* (2011) developed eighteen questions from theories of flow and play with responses measured using a five point Likert scale, where 5 is strongly agree and 1 is strongly disagree, matching the SUS scale. The questionnaire should be presented to users without groupings. For analysis, questions are categorised into seven groups derived from key elements of flow and play theory. *Challenge* captures the user's perception of how well the system balances boredom and anxiety. *Absorption* measures the potential for sustained use, including emotional investment and how well the games capture attention. *Appeal* measures how much a user likes the application. *Control* and *Learnability* relate to usability, control over actions and how quickly users learn to play the games. *Usefulness* is specific to GWAP and measures the users' perceived usefulness of the tagging system. Finally, *Social Interaction* measures how well the website supports competition, collaboration and communication between players.

A modified version of the questionnaire was used for the VideoTag evaluation. The social interaction category, which contained only one question, was removed because

VideoTag is not a social website and does not support interaction between players. A further eight questions were added to the questionnaire that pertain to the perceived usefulness of VideoTag and intention to continue to play, as well as questions that separately gauge user opinions of Golden Tag, Top Tag and Simply Tag. See Appendix B for the questions and Table 6-4 for the question groupings. Participants for the questionnaire⁸ were recruited following the methods outlined for the usability study in Section 6.2.

6.4.2 Results

A total of 7 participants answered all 25 questions. The mean, mode, median for all questions were calculated before the responses to negatively phrased questions were converted to a positive scale (Table 6-3). Most scores fell in the mid-positive range, with most means being between 3 and 4. There were few negative responses and some highly positive responses. The mode is most useful for gauging user agreement on the effect of design elements on the overall experience, but for questions with multiple modes the mean and median are required. There was only one negative result (mode<3); users needed to read the instructions before playing. Questions 7 and 25 were negatively phrased so the low mode is positive; users did not agree with the statements. Users did not find VideoTag to be difficult or stressful and they preferred to use the games than Simply Tag. Neutral results (mode=3) need to be compared with the position on the scale of the median and in particular, the mean. A positive mean indicates that more users agree with the statement and a negative mean indicates more users disagree with the statement. Questions 9 and 22 had a

⁸ <http://www.videotag.co.uk/playability.php>

mode of 3 with a negative mean (<3) indicating users did not find VideoTag intellectually stimulating and had little intention to continue to play it. Questions with a mode of 3 and a mean of 3 indicate users did not have a strong opinion in either direction for questions 2, 4, 6, 12, 13, 16 and 21. These questions related to system errors, enjoyment and level of challenge. No questions had a mode of 3 and a high mean (>4) indicating that there was more user agreement when a response was positive.

The mode and mean were compared for positive responses, revealing attributes of VideoTag that are strongest for playability. Highly positive responses (mode=5) with high means (≥ 4) were received for questions 4, 14, 17, 18 and 24. Users felt that they could learn to play the games quickly and get help if required. They found VideoTag to be a useful tool for tagging videos and the different goals of the two games were clear. Questions 1, 3, 5, 8, 10, 11, 20 and 23 had a high mode (4) of response scores and a mid-range mean (≤ 4) revealing that users felt VideoTag catered for players of different abilities and that it is sufficiently challenging; players' perceived skills match the perceived challenge and the games hold their attention. The majority of users felt that VideoTag is worth playing and they preferred Top Tag to Golden Tag. Users felt that VideoTag had a simple interface, they were encouraged to enter different words, were motivated by the time limit and felt their actions affected their scores. These findings are summarised in Table 6-4.

Table 6-3 The mean, mode and median of responses to the playability questionnaire (n=7).

Question	Question Description	Mean	Mode	Median
Q1	sufficiently challenging	3.7	4	5
Q2	challenge level	3.6	3	3
Q3	different skill levels	4.0	4	4
Q4	challenging over time	3.1	3	3
Q5	simple interface	3.7	4	4
Q6	felt bored playing	3.3	3	3
Q7	difficult and stressful	3.7	4	5
Q8	stay focussed	3.6	4	4
Q9	intellectually stimulating	2.9	3	2
Q10	motivated by time limit	3.7	4	3
Q11	actions impact on score	4.0	4	4
Q12	prevent errors	3.0	3	3
Q13	recover from errors	3.3	3	3
Q14	easy to learn	4.0	5	5
Q15	no need for instructions	2.9	2	2
Q16	learning to play	3.1	3	3
Q17	help available	4.0	5	5
Q18	useful tool	4.6	5	5
Q19	create keywords	3.9	4	5
Q20	worth playing.	3.6	4	4
Q21	enjoy playing	3.1	3	3
Q22	continue to play	2.6	3	3
Q23	prefer Top Tag	3.4	4	4
Q24	understood difference in gameplay	4.6	5	5
Q25	preferred Simply Tag	4.0	4	5

Table 6-4 Grouping structure of questions and summary of findings.

Category	Questions	Main Findings
Challenge	Q1 Q2 Q3 Q4	VideoTag is able to challenge people with different skill levels and is sufficiently challenging although challenge across levels could be improved.
Absorption	Q7 Q8 Q9 Q10	Users were motivated by the scoring system. Whilst VideoTag was not intellectually stimulating there was enough stimulation for users to keep focus and concentrate on the task with minimal anxiety.
Appeal	Q5 Q6 Q21 Q22 Q23	VideoTag had a simple and well designed interface. Users' showed a slight preference toward Top Tag over Golden Tag, although few users intended to play VideoTag again. The balance between anxiety and boredom could be improved which might encourage more users to continue to play.
Control	Q11 Q12 Q13	Users' were able to make errors but felt they could recover from them. They felt their actions in the games could impact their score.
Learnability	Q14 Q15 Q16 Q17 Q24	Users learned how to play the games quickly but needed to read the instructions first. Differences in game play were clear and help was available if required.
Usefulness	Q18 Q19 Q20 Q25	Users found VideoTag to be a useful tool that is worth playing, most felt the system encouraged them to enter new keywords. Users preferred to play games rather than use Simply Tag.

A Pearson's correlation analysis identified many significant positive and negative correlations between questions (see Table 6-5), most at the 95% confidence level $p \leq .05$ and some at the 99% confidence level $p \leq .01$. Questions 1, 9, 10, 13 and 23 had no significant correlations with any other question. Zero correlation was found between seven question pairs (2,11), (2,14), (2,17), (2,25), (3,8), (3,13), (4,11) exposing a weak relationship between Challenge and the other categories. The amount of significant correlations between questions indicates that further tests can be conducted to analyse relationships between variables and groups of variables.

Table 6-5 Pearson correlations for each question pair (n=7).

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25
Q1	1	.548	.382	.588	.367	-.047	.179	.028	.210	.507	-.127	.403	-.047	-.408	.193	.254	-.564	-.167	.574	.228	.319	.335	-.380	-.510	.630
Q2	.548	1	.837*	.867*	.428	-.389	-.130	-.091	.326	.482	.000	.529	.154	.000	.067	.435	.000	-.091	.080	.300	.548	.650	.344	-.062	.000
Q3	.382	.837*	1	.837*	.489	-.651	-.273	.000	.303	.367	.250	.791*	.000	.401	-.418	.607	.158	.382	.336	.627	.764*	.837*	.540	.259	-.177
Q4	.588	.867*	.837*	1	.208	-.247	.130	-.335	.266	.745	.000	.706	-.034	-.075	-.222	.242	-.265	.091	.295	.517	.517	.517	.459	-.227	.197
Q5	.367	.428	.489	.208	1	-.803*	-.838*	.782*	-.028	-.307	.651	.515	.359	.522	-.026	.918**	.206	.284	.406	.292	.533	.837*	.101	.531	-.345
Q6	-.047	-.389	-.651	-.247	-.803*	1	.851*	-.711	-.325	.171	-.814*	-.721	-.479	-.870*	.519	.960**	-.618	-.711	-.469	-.662	-.835*	-.934**	-.369	-.821*	.691
Q7	.179	-.130	-.273	.130	-.838*	.851*	1	-.863*	.024	.558	-.819*	-.345	-.535	-.802*	.174	-.852*	-.604	-.447	-.157	-.245	-.447	-.701	-.281	-.870*	.772*
Q8	.028	-.091	.000	-.335	.782*	-.711	-.863*	1	.132	-.520	.764*	.242	.609	.612	-.091	.795*	.483	.417	.403	.228	.417	.548	-.196	.679	-.540
Q9	.210	.326	.303	.266	-.028	-.325	.024	.132	1	.525	.152	.384	.558	.162	-.351	.289	.480	.364	.393	.580	.712	.326	-.172	.090	-.215
Q10	.507	.482	.367	.745	-.307	.171	.558	-.520	.525	1	-.183	.464	.090	-.392	-.234	-.191	-.348	.040	.335	.482	.320	.022	.113	-.544	.389
Q11	-.127	.000	.250	.000	.651	-.814*	-.819*	.764*	.152	-.183	1	.632	.644	.802*	-.558	.759*	.474	.764*	.504	.627	.573	.627	.360	.778*	-.707
Q12	.403	.529	.791*	.706	.515	-.721	-.345	.242	.384	.464	.632	1	.271	.507	-.706	.672	.100	.725	.744	.926**	.845*	.794*	.455	.328	-.224
Q13	-.047	.154	.000	-.034	.359	-.479	-.535	.609	.558	.090	.644	.271	1	.344	-.085	.484	.543	.281	.185	.333	.445	.333	.022	.414	-.607
Q14	-.408	.000	.401	-.075	.522	-.870*	-.802*	.612	.162	-.392	.802*	.507	.344	1	-.671	.730	.761*	.816*	.269	.559	.612	.671	.481	.971**	-.850*
Q15	.193	.067	-.418	-.222	-.026	.519	.174	-.091	-.351	-.234	-.558	-.706	-.085	-.671	1	-.326	-.353	-.943**	-.576	-.867	-.624	-.400	-.459	-.496	.394
Q16	.254	.435	.607	.242	.918**	-.960**	-.852*	.795*	.289	-.191	.759*	.672	.484	.730	-.326	1	.480	.563	.524	.562	.795*	.943**	.172	.697	-.536
Q17	-.564	.000	.158	-.265	.206	-.618	-.604	.483	.480	-.348	.474	.100	.543	.761*	-.353	.480	1	.483	-.106	.265	.483	.397	.228	.820*	-.894**
Q18	-.167	-.091	.382	.091	.284	-.711	-.447	.417	.364	.040	.764*	.725	.281	.816*	-.943**	.563	.483	1	.660	.867*	.708	.548	.354	.679	-.540
Q19	.574	.080	.336	.295	.406	-.469	-.157	.403	.393	.335	.504	.744	.185	.269	-.576	.524	-.106	.660	1	.783*	.660	.502	-.190	.100	.119
Q20	.228	.300	.627	.517	.292	-.662	-.245	.228	.580	.482	.627	.926**	.333	.559	-.867*	.562	.265	.867*	.783*	1	.867*	.650	.344	.372	-.296
Q21	.319	.548	.764*	.517	.533	-.835*	-.447	.417	.712	.320	.573	.845*	.445	.612	-.624	.795*	.483	.708	.660	.867*	1	.867*	.216	.481	-.405
Q22	.335	.650	.837*	.517	.837*	-.934**	-.701	.548	.326	.022	.627	.794*	.333	.671	-.400	.943**	.397	.548	.502	.650	.867*	1	.344	.589	-.444
Q23	-.380	.344	.540	.459	.101	-.369	-.281	-.196	-.172	.113	.360	.455	.022	.481	-.459	.172	.228	.354	-.190	.344	.216	.344	1	.427	-.509
Q24	-.510	-.062	.259	-.227	.531	-.821*	-.870*	.679	.090	-.544	.778*	.328	.414	.971**	-.496	.697	.820*	.679	.100	.372	.481	.589	.427	1	-.917**
Q25	.630	.000	-.177	.197	-.345	.691	.772*	-.540	-.215	.389	-.707	-.224	-.607	-.850*	.394	-.536	-.894**	-.540	.119	-.296	-.405	-.444	-.509	-.917**	1

*. Correlation is significant at the 0.05 level (2-tailed).#. Correlation is significant at the 0.01 level (2-tailed).@

H0 - The playability questionnaire is not a reliable measure of playability.

To conduct the reliability analysis, responses to negatively phrased questions (6, 7 and 25) were re-coded from negative to positive. The Cronbach's alpha indicated

high reliability ($\alpha=.90$) and redundancy was found in questions 1, 10 and 15. The null hypothesis can be rejected so the scale is a reliable measure of playability and confidence can be placed in the results.

H0 - Perceived usefulness does not affect user intention to play.

The descriptive statistics gave evidence that users perceived VideoTag to be useful and questions relating to usefulness in the playability questionnaire had the highest means. Hsu and Lu (2004) and Shin and Shin (2011) attribute perceived usefulness to intention to play games. To explore this further, paired t-tests were conducted to see whether perceived usefulness had an effect on users' intention to play, measured by Appeal and Absorption. (Appeal + Usefulness) and (Absorption + Usefulness) had significant relationships at the 95% confidence level ($p=.034$ and $p=.036$, respectively). There were no other significant relationships (see Table 6-6). There was some evidence that perceived usefulness had more impact on user enjoyment of VideoTag than on perceptions of challenge or ease of use. The VideoTag playability evaluation supports Goh et al.'s (2011) findings that users prefer to tag in a game rather than a non-game environment. This research extends the work of Goh *et al.* (2011) by analysing relationships between individual question categories and the influence certain game characteristics have on motivation to use a gamified tagging system.

Table 6-6 Paired t-tests for differences in mean between question categories (significant results are highlighted in grey) (n=7).

Pairs	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Challenge - Appeal	.42857	.36886	.18443	-.15836	1.01550	2.324	3	.103
Challenge - Absorption	.14286	.71903	.35952	-1.00128	1.28700	.397	3	.718
Challenge - Usefulness	-.39286	.61029	.30514	-1.36396	.57824	-1.287	3	.288
Challenge - Control	.33333	.54085	.31226	-1.01021	1.67688	1.067	2	.398
Challenge - Learnability	.10714	.81962	.40981	-1.19705	1.41133	.261	3	.811
Appeal - Absorption	-.28571	.61721	.30861	-1.26784	.69641	-.926	3	.423
Appeal - Usefulness	-.82143	.44224	.22112	-1.52513	-.11772	-3.715	3	.034
Appeal - Control	-.04762	.29738	.17169	-.78635	.69112	-.277	2	.808
Appeal - Learnability	-.48571	.77985	.34876	-1.45402	.48260	-1.393	4	.236
Absorption - Usefulness	-.53571	.29451	.14725	-1.00434	-.06709	-3.638	3	.036
Absorption - Control	-.04762	.54085	.31226	-1.39116	1.29592	-.152	2	.893
Absorption - Learnability	-.03571	.50000	.25000	-.83133	.75990	-.143	3	.895
Usefulness - Control	.57143	.28571	.16496	-.13833	1.28118	3.464	2	.074
Usefulness - Learnability	.50000	.41239	.20620	-.15621	1.15621	2.425	3	.094
Control - Learnability	.09524	.08248	.04762	-.10965	.30013	2.000	2	.184

These findings, coupled with the MPRC findings, support research by Hsu and Lu (2004) and Shin and Shin (2011) that perceived usefulness and ease of use of the system influences engagement and intention to play over perception of challenge and goals. The findings of the playability study found that challenge did not correlate with appeal or absorption, contrasting with findings by Febretti and Garzotto (2009), that challenge was highly correlated with engagement in games. This indicates that users perceived VideoTag as being a tagging system more than a game. Therefore, rules that apply to why people play games cannot easily be applied to VideoTag games. Users will be attracted by the perceived usefulness and ease of use of the system and not by the challenges posed by the games. However, Chapter 5 revealed that most users who used VideoTag played many games, and played more games than they used Simply Tag. This indicates that although not perceived as a standalone game, the game elements motivated use and helped to engage users. Regardless of how many game elements are applied to the system, users will not play for the sole purpose of being entertained. They will use the system because they

have an interest in the content and perceive the usefulness of tagging the videos, but they will use the system more intently if game elements are present.

6.5 Engagement and Enjoyment

To investigate the findings of the user studies further, an evaluation was conducted of each experimental phase of VideoTag and the individual games using the Sweetser and Wyeth (2005) Game Flow model. The enjoyment evaluation focuses on the integration of game elements to gain an understanding into what aspects of the aesthetics and mechanics of the game could motivate or demotivate users and the likelihood of users enjoying the games. In light of the poor user participation reported in Chapter 5 and the lack of enjoyment highlighted in the playability study, the focus will be on potential reasons why users were deterred and what elements could be improved in the future.

For the Game Flow model, Sweetser and Wyeth (2005) identify eight core elements that effect the enjoyment of games: *Concentration, Challenge, Skills, Control, Clear Goals, Feedback, Immersion* and *Social Interaction*. These eight elements form the core structure of their Game Flow game enjoyment model with heuristics applied to each of the eight elements. See Table 6-7 for the evaluation scheme and each heuristic. An outline of what each core element of the model evaluates is discussed below:

Concentration - The more a user can concentrate on a game the more enjoyable it will be. A game will be more absorbing if it requires a large workload and more of the user's attention to complete tasks. When a user needs to use all of their skills to cope

with in-game challenges, they are unaware of external influences. The interface of the game needs to grab a user's attention quickly and maintain it throughout the game, however long the game lasts. Game action should dominate the screen; there should be no visible or audible distractions to interrupt a user's focus on the game itself. For example, in-game adverts massively detract user attention and interrupt gameplay, creating frustration that can reduce user enjoyment.

Challenge - The most important aspect of good game design is challenge and, as Febretti and Garzotto (2009) discovered, it has a high correlation with user engagement. Games should match a user's skill level and difficulty should be varied. Games should keep an appropriate pace increasing in difficulty as the user gets more familiar with the game. If a game is too difficult then this creates anxiety but if it is too easy then this creates apathy. The rewards of challenge are intrinsic, indicating that the correct level of challenge has a high probability of initiating flow.

Player Skills - The game should support users in developing and mastering skills to progress through levels. There should be scope for the game to get more difficult as a user's skills improve. Users should be able to learn whilst playing without the need to read a manual or list of instructions (Desurvire et al., 2004) although help should be available during the game if a user requests it.

Clear Goals - All games need goals to motivate users to play. They should be presented to the user early in the game. Each level should have multiple goals and

obstacles. To engage a user and to achieve flow, goals must be clear (Csikszentmihalyi, 1975).

Control and Feedback - The criterion for Control and Feedback are based strongly in usability heuristics rather than flow, although both are a pre-requisite for flow. Users need to feel in control of their own actions and not controlled by the game. Users should be able to leave a game when they want to by pausing or saving it. Errors can make users feel that they have lost control, but Febretti and Garzotto (2009) found that most users would find a way to overcome or bypass usability errors and that they did not deter a user from continuing to play. Users should not be able to make mistakes that stop the game from working. They should be able to recognise, diagnose and recover from errors. Users should feel that their actions and decisions have an impact on the game that affects the experience. The game should allow the user the choice to play the game how they want, creating multiple paths to play the same game, essentially making the game different every time. Players must receive appropriate feedback at appropriate times. Concentration is possible when the task provides immediate feedback (Csikszentmihalyi and LeFevre, 1989). A game should provide frequent in-game feedback so users can determine their progression toward objectives. The interface and sound can be used to give status feedback (Juul, 2009).

Immersion - is the flow state; it is the desire to devote extreme amounts of time, effort and attention into playing a game. An immersed user will have a high emotional investment in the game. Flow is difficult to measure and observe, criterion can be defined that could result in flow, but it is not guaranteed.

Social Interaction - Social Interaction encompasses virtual community, chat and competition between friends. Cowley *et al.* (2008) query the inclusion of Social Interaction in the Game Flow model questioning whether it is a necessary or desirable aspect of every game. Many researchers (Lazzaro, 2008; Hunicke *et al.*, 2004; Bartle, 1996; Cowley *et al.*, 2008; Bateman and Boon, 2006; Yee, 2006) include social interaction when modelling user behaviour providing the evidence to include it. Sweetser and Wyeth (2005) surmise that social interaction is not an element of flow and can interrupt the flow state, yet many people play games for the social interaction they provide.

6.5.1 Methods

The Game Flow model (Sweetser and Wyeth, 2005) was used to evaluate video tagging games for potential user enjoyment. Since an independent evaluator was not used the results may be biased. Four evaluations were conducted: The VideoTag phase one website, each individual game, Golden Tag and Top Tag and the phase two website. Only the content of the videos and the VideoTag website changed in phase two. The games did not change and so were only evaluated once. The website was evaluated because game elements are present on the website as well as in the individual games. Each heuristic is measured using a five-point scale: 0 – N/A, 1 – not at all, 2 – below average, 3 – average, 4 – above average, 5 – well done. Average scores for each section were calculated followed by the overall average and percentage score.

6.5.2 Results

Table 6-7 Results of the four enjoyment evaluations using the Game Flow model (Sweetser and Wyeth, 2005).

	Criteria	VT-p1	GT	TT	VT-p2
Concentration Games should require concentration and the player should be able to concentrate on the game	• games should provide a lot of stimuli from different sources	4	4	5	5
	• games must provide stimuli that are worth attending to	4	5	5	5
	• games should quickly grab the player's attention and maintain their focus throughout the game	4	5	5	5
	• players shouldn't be burdened with tasks that don't feel important	4	5	5	3
	• games should have a high workload, while still being appropriate for the players' perceptual, cognitive, and memory limits	0	5	5	5
	• players should not be distracted from tasks that they want or need to concentrate on	4	5	5	4
	Total		4	4.8	5
Challenge Games should be sufficiently challenging and match the player's skill level	• challenges in games must match the players' skill levels	5	5	5	5
	• games should provide different levels of challenge for different players	1	3	3	5
	• the level of challenge should increase as the player progresses through the game and increases their skill level	1	2	2	2
	• games should provide new challenges at an appropriate pace	1	1	1	3
Total		2	2.75	2.75	3.8

Player Skills Games must support player skill development and mastery	<ul style="list-style-type: none"> players should be able to start playing the game without reading the manual 	5	5	5	5
	<ul style="list-style-type: none"> learning the game should not be boring, but be part of the fun 	1	5	5	5
	<ul style="list-style-type: none"> games should include online help so players don't need to exit the game 	2	5	5	5
	<ul style="list-style-type: none"> players should be taught to play the game through tutorials or initial levels that feel like playing the game 	1	1	1	5
	<ul style="list-style-type: none"> games should increase the players' skills at an appropriate pace as they progress through the game 	0	3	3	3
	<ul style="list-style-type: none"> players should be rewarded appropriately for their effort and skill development 	3	4	4	5
	<ul style="list-style-type: none"> game interfaces and mechanics should be easy to learn and use 	0	5	5	5
	Total		2.4	4	4
Control Players should feel a sense of control over their actions in the game	<ul style="list-style-type: none"> players should feel a sense of control over the game interface and input devices 	4	4	4	5
	<ul style="list-style-type: none"> players should feel a sense of control over the game shell (starting, stopping, saving, etc.) 	3	3	3	3
	<ul style="list-style-type: none"> players should not be able to make errors that are detrimental to the game and should be supported in recovering from errors 	4	2	2	5
	<ul style="list-style-type: none"> players should feel a sense of control and impact onto the game world (like their actions matter and they are shaping the game world) 	2	1	1	4
	<ul style="list-style-type: none"> players should feel a sense of control over the actions that they take and the strategies that they use and that they are free to play the game the way that they want (not simply discovering actions and strategies planned by the game developers) 	4	4	4	4
Total		3.4	2.8	2.8	4.2

Clear Goals Games should provide the player with clear goals at appropriate times	<ul style="list-style-type: none"> over-riding goals should be clear and presented 	2	2	2	5
	<ul style="list-style-type: none"> early - intermediate goals should be clear and presented at appropriate times 	2	2	2	5
	Total	2	2	2	5
Feedback Players must receive appropriate feedback at appropriate times	<ul style="list-style-type: none"> players should receive feedback on progress toward their goals 	2	3	4	4
	<ul style="list-style-type: none"> players should receive immediate feedback on their actions 	2	5	5	5
	<ul style="list-style-type: none"> players should always know their status or score 	5	3	4	5
	Total	3	3.7	4.3	4.7
Immersion Players should experience deep but effortless involvement in the game	<ul style="list-style-type: none"> players should become less aware of their surroundings 	0	1	1	1
	<ul style="list-style-type: none"> players should become less self-aware and less worried about everyday life or self 		1	1	1
	<ul style="list-style-type: none"> players should experience an altered sense of time 		2	2	1
	<ul style="list-style-type: none"> players should feel emotionally involved in the game 		2	2	3
	<ul style="list-style-type: none"> players should feel viscerally involved in the game 		1	1	2
	Total		1.4	1.4	1.6
Social Interaction Games should support and create opportunities for social interaction	<ul style="list-style-type: none"> games should support competition and cooperation between players 	3	1	1	3
	<ul style="list-style-type: none"> games should support social interaction between players (chat, etc.) 	1	1	1	1
	<ul style="list-style-type: none"> games should support social communities inside and outside the game 	1	1	1	3
	Total	1.7	1	1	2.3
	Overall	2.6	2.8	2.9	3.85
	Overall %	52%	56%	58%	77%

Key - VT-p1=VideoTag website phase one; GT=Golden Tag; TT=Top Tag; VT-p2=VideoTag website phase two.

The results confirm that design changes in phase two improved the overall experience of using VideoTag, with an almost 20% improvement. Major improvements were in Player Skills and Clear Goals sections, with users being better supported to learn how to use the system and identify its purpose in phase two. Scores for Feedback and sense of competition and collaboration increased with improved integration of the level thermometer and allowing users to experience how the tags they enter could be used with search and browse provision. Results of the evaluation will be discussed in more detail with reference to findings from the user studies in the next section.

6.5.3 Discussion

The ability to concentrate on given tasks and maintain focus is a prerequisite for enjoyment. Phase two of VideoTag was designed to attract the user's attention immediately to the step-by-step tasks to begin tagging videos. Whilst there are many graphical stimuli on the interface, they are all relevant to the task. Respondents felt VideoTag had a simple and well designed interface which is intuitive, understandable, functional, clear and fit for purpose. A high cognitive workload for the games was recorded as a problem in the evaluation. It takes skill and concentration to watch a video, interpret the content, think of relevant tags and type them into the system, coupled with additional stimuli from gameplay elements such as finding the golden tag, or finding the top tags, all whilst under the pressure of time. VideoTag requires concentration and the use of many skills and so, theoretically, it should be an absorbing task. Informal observations reported in Chapter 5 suggest that some users were not absorbed in the activity and the skill involved could be a deterrent. However, a majority of respondents did not find the games difficult or stressful and were able to maintain focus and concentrate on

tagging videos without unnecessary distractions. The results reported in Chapter 5 suggest that VideoTag did not appeal to many users because it failed to attract the attention of many visitors who followed direct links. However, of the participants who did play, many users played many games. This suggests that attention was maintained, although not for a sustained period as most playability questionnaires expressed no interest in continuing to play VideoTag. This could be attributed to the prototype form of VideoTag. It is not an established video sharing system with a community of users, content is not changing and there is no long term benefit to the user from tagging the videos.

Challenges match player skills and have the potential to appeal to different users. This is supported by the playability evaluation. Challenge does not increase over time or over different levels although only a few participants in the playability study recognised this as a problem. Levels are poorly defined, existing primarily as a rank on the leaderboard. Level progression in Golden Tag is indicated by reaching a new decade; users are rewarded with a change of interface and a new selection of videos, but not with new challenges. Goals of the games were clearly presented in phase two and users were clear about the differences between gameplay in both games. More goals for players are needed to increase curiosity about what the next level will hold, motivating users to continue. Designing levels is difficult when the primary function of the game is to tag videos with little existing textual data. Any goals should be relevant to the content and the activity. Golden Tag supports the challenge to enter tags of higher quality but this behaviour is likely to go unrewarded unless another user has already entered the tag and probability of agreeing on subjective vocabulary is low. Higher levels could utilise the tag data generated in lower levels acting as

both a motivator and also reinforcing that the data users enter is useful and has purpose to the game experience as well as the whole VideoTag experience.

Descriptions of what tagging is, the purpose of the games and how to play were inadequate in phase one of the VideoTag website. The number of clicks needed to begin playing was minimised but the design assumed that all visitors would know what was required of them. This was rectified in phase two; users were given graphical step-by-step instructions on how to begin to play as well as information boxes pertaining to purpose. This improvement in phase two is supported by the user studies. Users needed to read the instructions before they could play but the majority felt the interface was easy to use and learn. Learning was not part of the fun and this could be improved by incorporating learning into initial rounds of the games, taking players through example rounds and initially suggesting suitable tags. Users experience an initial increase in skills and then a plateau and so there is no increase in difficulty for their skills to continue to improve. The game is easy to learn but a lack of scope to increase skills creates boredom and will not maintain use. Users feel adequately rewarded for their skills and feel that their actions in the game impact their scores. In phase two the leaderboard was presented at the end of each game so that it was more integrated with the actions of the games. Users could see instantly if they had progressed to a new level. However, as discussed previously, level progression did not bring enough rewards to sustain play and encourage future use. Top Tag supports feedback better than Golden Tag as users are continuously aware of their level and progress even though there is only a distinguishable difference between levels in Golden Tag.

Potential errors highlighted in the enjoyment evaluation are the result of the games being browser based and built as web applications rather than downloadable games. Server load is the error most likely to affect users. Other errors are most likely to be replicated by inexperienced web users or users trying to break the system. The user studies rated VideoTag as having good usability; users described the system positively in relation to ease of use, indicating few usability problems. If users make errors they feel supported to recover from them. Users felt a sense of control through the provision of choices. Phase one allowed a choice of video category and game and phase two extended this to allow more choice over content. The option to have more control over the system by uploading content is awarded to users who invest time in the system. Usage statistics show that few users invested time in the system or engaged fully with all of its features. It was difficult to attract users freely out of interest in the content. Many features were dependant on a community of users and without this users had a reduced experience. Despite this, users felt that their actions could have impact on the system, that VideoTag is worthwhile and that it encourages them to enter tags.

Social interaction was not evaluated in the user studies. Design decisions for not making VideoTag a social game are discussed in Chapter 5. It was beyond the scope of this research to create a social layer within the VideoTag system although attempts were made to accommodate integration with existing social networks. Community is difficult to build, especially in a short time period with little promotion, and without backing from a trusted organisation with an existing good reputation. Shin and Shin (2011) report that perceived security is significantly related to a person's intention to play a game. VideoTag would benefit from a community of users from specific interest groups uploading and tagging video content. Without this, the potential

appeal of many features of the system is reduced. There is a basic provision for competition and cooperation in the form of the leaderboard and forming special interest groups. There is no provision for social interaction except for links to share on social networks and limited provision for social communities. Provision of a social layer would improve the overall enjoyment of VideoTag. Enjoyment might also be improved if VideoTag were deployed as a tool within existing communities.

Immersion is difficult to evaluate because it is internal to a user. If users are immersed then they are more likely to enter a flow state. There is no evidence in usage statistics or user studies that any VideoTag users entered a flow state, due to the lack of continued use. Users rated VideoTag higher for ease of use and usefulness as a system rather than for enjoyment. Some users said they did not enjoy VideoTag, most agreed they would not play again and some felt bored whilst playing. This evaluation has revealed areas of VideoTag that could be improved to motivate continued use, improve user enjoyment and potential for immersion. VideoTag suffers from a lack of community, lack of support for improving skills, a lack of levels and minimal variety in challenges. More levels should be integrated in to Golden Tag and Top Tag, increasing difficulty with higher levels so that players can improve their skills and match their improved skills to new appropriate challenges. More whole site goals and achievements could be implemented to improve users' perceptions that their contribution is worthwhile. Competition in VideoTag is supported by the level thermometer; whilst users feel rewarded for actions by the scoring system the evaluation suggests they are not motivated to compete to score more points and reach higher levels. An improvement in feedback and creating an achievements structure will support users motivated by competition. Only with community will users collaborate to reach a shared goal. Collaboration was supported in phase two with the addition of special interest categories, rather than the generic YouTube categories included in phase one, but without a community of

users, the features were not used to their full potential. Community is difficult to create and video tagging games need to be attributed to an existing community of users who will be motivated by altruism for the community and an interest in the content. The success of the Waisda? video tagging game (Hildebrand et al., 2013) that ran in parallel to VideoTag supports the notion that reputation of the organisation and an existing community of users affects the success of a video tagging game more than the quality of game. Perceived usefulness of the system overrides perceptions of enjoyment of the game experience.

6.6 Barriers to Use or Play

Disappointing user numbers are not explained by these evaluations. The user studies and enjoyment evaluation indicate that VideoTag provides a reasonably positive experience. Promoting VideoTag as a system and improving the perception of usefulness could attract more users than promoting the system as a suite of games with a purpose. A useful addition to the user studies would have been to question how users perceive the system to be useful and to gather more feedback on their opinions of content and how that affects their intention to use the system. This would be especially interesting based on the findings in Chapter 5 that content had an effect on system choice and the tags entered.

To capture a contrasting viewpoint and highlight any barriers to use and/or play, a questionnaire asking why users did not play was posted on the VideoTag Facebook page and personal social media sites of the author. Questions were formulated based on informal observations during the design and implementation process, discussed

in Chapter 5. The survey was informal, conducted as a method to qualify the informal observations. The following questions were devised:

Please select as many of the following reasons that apply:

1. I don't know what tagging is;
2. It didn't work on my phone or tablet;
3. I didn't know what I was supposed to do;
4. It looked complicated;
5. It looked like it would be too time consuming;
6. Didn't interest me;
7. There were no videos I liked;
8. I didn't want to sign up for an account;
9. I didn't trust it;
10. Other reason / more feedback.

Use of this questionnaire was informal and had six respondents, in line with the other user studies. Two users stated that it didn't interest them, two didn't want to sign up for an account, two did not know what tagging is, two thought it would be too time consuming and one said it was because it wouldn't work on their phone or tablet. The study is too minimal to quantify *amotivation* state; it reinforces suggestions that users are deterred by a lack of understanding and perceived usefulness.

The study highlights potential barriers to use. The playability study revealed that users had to read the instructions before they could play the games. This coupled with signing up for an account takes time. Whether users want to invest the time depends on their perception of how useful the system is. A lack of understanding of

what tagging is and why it is useful will negatively affect a user's motivation to use VideoTag. If they do not understand the purpose they will have a low perception of usefulness, struggle to interpret the goals and have a negative perception of their skills opposed to the challenge. The playability study showed perceived usefulness correlates to appeal, so if perceived usefulness is low, the system will have little appeal. The MPRC study highlighted ease of use as important to users of VideoTag. If players feel they have to learn too much before playing and then sign up for an account then the website is not easy to use. Users with a 'just try it out' motivation will be deterred because of the time investment involved before they can try it out. VideoTag does not have enough challenges or immersive experiences for hardcore players and was not designed for such players. The objective was to attract casual game players, harnessing the time they invest on quick and easy time filling games. Casual game players will be deterred because they want a game that is quick to play, easy to learn, takes little cognitive effort and balances frustration and boredom effectively. A casual player has no interest in overcoming frustration, but if not challenged enough will easily become bored and move on to another game. The evaluation revealed flaws in the current implementation of VideoTag, lack of challenge, goals, levels, reward and community. In its current form it does not offer enough of a gameful experience to most game players, it does offer a gameful experience to users who perceive purpose in tagging the videos.

A knowledge and understanding of tagging and its purpose as well as an interest in the video content are pre-requisites to the use of video tagging games. Without investment in the outcome of their interactions, users may be deterred by the time investment required to play the games. Despite the efforts to design a system layer in phase two that showed how the tags could be used, this was not explored by users

and the features were not fully engaged with. This can be attributed to a lack of community and a lack of interest in the content, despite efforts to attract special interest groups. Users will perhaps tag videos if they are interested in the content, want to share it with other users or organise it for their own collections. YouTube offers every service VideoTag needed to create, except social tagging. To create the social tagging outside of the YouTube community does not improve the YouTube experience for YouTube users. There is no perceived usefulness in tagging the videos, therefore there is no incentive to use the system. To attract users to use VideoTag as a tool to label online videos, collaboration with a trusted organisation with an existing community of users is required.

7 Classification Studies – Game Based Tagging of User Generated Video

7.1 Introduction

The Literature Review revealed a gap in knowledge about tagging GWAPs in that few projects have analysed the quality of their outputs. For the Waisda? video tagging project, Lin and Aroyo (2012) extend GWAP theory to encourage users to assess how accurately a tag describes the video. The proposed system aims to validate individual tags by assessing their quality, their meaning, their suitability for describing fragments of video and the suitability of the tag set at describing the whole video. Unfortunately, the literature only describes the system and as yet no results have been published. Gligorov *et al.* (2011) compared user tags entered in the Waisda? project to professional annotations. They found that tags identify objects in the video rather than interpreting topics or scenes, whereas professional annotations interpret the entire video. User tags can complement professional annotations, but few user tags matched the controlled vocabulary. Gligorov *et al.* (2013) suggest that user tags generated in the Waisda? video tagging project can improve video search over using professional annotations. They found that verified tags (i.e., tags with

high user agreement or matching tags) had high precision but lower recall. They found a search engine indexing all user tags performed 33% better than a search engine only indexing verified tags.

The majority of tagging projects utilise the Von Ahn (2006) model and assume without proof that high user agreement equates with high quality. This is not necessarily true from the perspective of information retrieval. If users tend to agree on basic terms then tags of higher semantic level may be rejected even though these are particularly useful for video search. Moreover, computer processing of visual content cannot yet extract keywords that describe objects at a subordinate level or subjective terms. The literature review highlighted that whilst there has been progress in classifying image tags from tagging systems like Flickr, few such studies exist for videos.

The most widely adapted model for interpreting image content is from Panofsky (1970) and Shatford (1986). Panofsky applies theory of iconography to interpret image content, defining three levels of meaning *pre-iconographic*, *iconographic* and *iconological*. The iconographic levels (*pre-iconographic* and *iconographic*) correspond to basic level theory (Rosch, 1975) which categorises the specificity at which text describes an object (see Section 2.6.1, p79). Panofsky's first level, *pre-iconographic*, categorises the recognition of objects in the image and factual information relating to the image at their most general, objective level. No subject specific knowledge is required to identify objects at the pre-iconographic level and descriptions will be easily identifiable and simplistic. This encompasses superordinate level, the most simplified description of an object (e.g., animal, place, building) and basic level, the

most easily identifiable descriptions of an object (e.g., dog, town, house). Panofsky's second level, *iconographic*, categorises the recognition of objects with more familiarity. Descriptions will be at a subordinate level, applying accurate knowledge of a subject at a specific, objective level (e.g., poodle, Ambleside, 'dunroamin' as the name of the house). Panofsky's third level, *iconological*, categorises interpretation of what the objects symbolise and what the whole image means at an abstract and subjective level (e.g., retirement, tranquillity, content). Shatford (1986) identifies subject specific attributes that describe what an image is 'of' or 'about'. What an image is 'of' is concrete and objective, whereas what an image is 'about' is abstract and subjective. Each attribute describes the 'who', 'what', 'where' and 'when' of an image (e.g., Who created the image? What was the creator capturing? Where was it taken and when?). What an image is 'of' encompasses Panofsky's pre-iconographic and iconographic levels and basic level theory. The 'of' attribute identifies objects in an image at either a basic (pre-iconographic) level (e.g., city, park, dog), or at a specific (iconographic) level (London, Clapham Common, Pug). The 'about' attribute interprets the visual content at an abstract and subjective level, using the knowledge used to identify what the image is 'of' an idea of what it is 'about' is formed. This encompasses the Panofsky iconological level (e.g., fun, enjoyment, lazy, relaxed).

Jaimes and Chang (2000) extend the Panofsky (1970) / Shatford (1986) model to incorporate syntax information as well as semantic information. Their ten level model for interpreting visual content incorporates non-visual, perceptual and conceptual descriptions. Non-visual elements include date, location and creator; these elements are available as textual data and can be indexed easily. Elements at the perceptual level include colour, shape and texture. These low level features can be extracted by image-processing algorithms for content based search. At the

conceptual level in the extended model, visual content is interpreted using six attributes with objects and scenes described at a basic, specific and abstract level. Rafferty and Hilderley (2007) propose a scheme for classifying image tags using the Panofsky/Shatford model. Similarly to Jaimes and Chang (2000) the scheme models perceptual level features as well as conceptual. Three levels of meaning are described based on Panofsky (1970) with each level identifying tags that describe the 'who, where, what and when' attributes defined by Shatford (1986). Both factual and interpretive qualities of the tags are classified. The first level does not index image meaning but captures perceptual level elements of the image. The second level records specific descriptions of objects and more specific facts about the image. The third level is subjective and captures tags that interpret the whole image to incorporate expressions of opinion.

Jørgensen *et al.* (2001) evaluated the ability of the Jaimes and Chang (2000) classification scheme to classify descriptions of images created by naive users and professional indexers. They found that 87% of descriptions were conceptual. More objects (70.3%) were described than scenes (29.7%) at a basic rather than a specific or abstract level. Hollink *et al.* (2004) used the model defined by Jaimes and Chang (2000) to classify user descriptions of a set of images; they found that most users preferred to use general descriptions at the basic level than to provide specific or abstract interpretations; a finding supported by Ransom and Rafferty (2011). Gligorov *et al.* (2011) classified 1354 tags entered during a pilot study of the Waisda? video tagging project using a variation of the Panofsky (1970) / Shatford (1986) and Jaimes and Chang (2000) models. In line with the findings of Jørgensen *et al.* (2001), 1343 conceptual tags were found with only 11 perceptual tags and no non-visual tags. Similarly to Jørgensen *et al.* (2001) and Hollink *et al.* (2004), 74% of tags were general, 9% specific and 7% abstract. Kim (2011) classified the tags of 300 YouTube videos using a custom scheme derived from the Panofsky (1970) / Shatford (1986)

and Jaimes and Chang (2000) models. Most descriptions were conceptual (51.7%) at a basic or specific level followed by abstract descriptions (29.9%) although, the highest attribute of this category refers to basic level descriptions of YouTube categories (21.6%) which in the Angus et al. (2008) classification scheme (see Table 7-1) would be an A2 basic objective tag and not classified as an abstract description. Similarly to Gligorov *et al.* (2011), Kim (2011) found few tags at the non visual or perceptual levels. Kim (2011) identified that tags assigned to videos were closer to professional indexer assigned terms than tags assigned to images as more tags describe the video content. In contrast Gligorov *et al.* (2010) found user tags had low agreement with professional annotations. Both Kim (2011) and Gligorov *et al.* (2013) state that tags assigned to videos are useful as additional metadata to improve video search. Kim (2011) claims that further semantic expansion of existing YouTube tag data is required. The findings of Kim (2011) and Gligorov *et al.* (2011) suggest that the Jaimes and Chang (2000) model is not suited to classifying user generated tags that are predominately conceptual. It is best suited to the classification of systems using automatic methods to annotate videos, professional annotations or a combination of these methods with user tags. The classification scheme used in Chapter 4, adapted from Angus *et al.* (2008) is more appropriate for tag classification because it classifies the descriptive vocabulary of conceptual level tags.

Classification of the language used in video tags may help to indicate whether tags can be used to index video and improve video search. Tagging is useful at bridging the gap between automated methods and the natural language of queries. Gligorov *et al.* (2011) question whether users can be encouraged to use more descriptive vocabulary. In a time pressured environment it is easier to think of a basic level tag than a specific or subjective one (Gligorov et al., 2010). Games may produce more

basic level tags of more objective vocabulary than Simply Tag because they are easier to think of in the time constrained tagging environment. More misspellings may also be generated in games than the non-game system. However, as tagging is prolonged in the game environment is it more likely that users will start entering basic level tags and then progress to more specific tags (Goh et al., 2010) as with image tagging. Gligorov *et al.* (2010) suggest that the temporal nature of video affects how users tag. Results from the preliminary studies (Chapter 4) show that users tag video with more specific language than for images. During the design and implementation process of how aspects of game play and video choice might affect the types of tag users enter. Hunting for the Golden Tag (a tag entered by only one other player) might encourage users to enter more specific objective tags and more subjective tags; also asking users to find the Top Five tags for a video might encourage them to enter more basic level tags. The results outlined in Chapter 5 confirmed that Top Tag generated tags with the highest levels of tag agreement, indicating that gameplay affected the specificity of tags, with tags of higher specificity being entered into Golden Tag. The results also suggested that users enter more basic level tags into a game environment and that selecting videos with specific categories with the aim of attracting users from special interest groups generated more specific level tags. This chapter assesses whether game elements and content have an effect on the types of tag that users enter, investigating the level at which the tags describe or interpret the video content. Through a classification of VideoTag tags similarities in how users tag video in a game and non-game environment will be evaluated. The tag classification will investigate further whether gameplay affected the types of tag users entered.

7.2 Methods

7.3 The Classification Scheme

Jaimes and Chang (2000), Kim (2011) and Gligorov *et al.* (2011) found few instances of tags that record non visual features or perceptual features, but the vast majority of tags described conceptual elements of video content. The preliminary studies discussed in Chapter 4 found that a classification scheme adapted from Angus *et al.* (2008) was successful at highlighting tagging practice on YouTube and Viddler and assessing the descriptive quality of the tags assigned to videos on each system. The scheme (see Table 7-1) extends the Panofsky (1970) / Shatford (1986) model to include categories specific to tagging practice, based on the tag functions defined by Golder and Huberman (2006). Ransom and Rafferty (2011) claim that the Panofsky/Shatford model on its own is insufficient at classifying tags because 22% of tags in their dataset fell outside of the Shatford matrix. Sections A and B of the Angus *et al.* (2008) scheme relate closely to the Panofsky (1970) / Shatford (1986) model, detailing levels of description and abstraction. Sections C and D closely model the tag functions defined by Golder and Huberman (2006) and allows for evidence of the vocabulary problem (Furnas *et al.*, 1987) to be classified. Based on the findings of the preliminary studies a few amendments were made to the scheme:

- A specific objective category (B1c) was added for tags that describe specific actions.
- The most frequently used tag type on YouTube was D2 - multi-word tags; D2 tags were also prevalent on Viddler. The VideoTag system was designed to accept multi-word tags. A specific objective category (B1d) was created for multi-word tags that explicitly describe the video content and create additional textual data that includes people's names, phrases and whole sentences.

- The irrelevant multi-word category (D2) remains to categorise multi-word tags that repeat textual data already available, such as video title or other irrelevant multi-words.
- Despite few instances of URLs in the YouTube or Viddler datasets D9 was left in to record spam.

Table 7-1 The classification scheme detailing each tag type category

	A	Generic relationship between tag and video content		
<i>Pre- Iconographic</i>	Tag identifies what the video is of at its most primary and objective level - no subject specific knowledge is needed to make this distinction.		Basic	Objective
	1	Tag generically identifies what video is 'of'. e.g. a video of a cat, tagged as 'cat' or 'animal', characters tagged as 'man' or 'kid' or objects e.g. 'guitar', 'car', 'football'. Tags also identify things that are easily identifiable in visuals or audio e.g. a person's name or location said multiple times, written on screen or appear in the title of the video . Also included is the tag 'video'.		
	2	Tag identifies video Category/Genre. General YouTube defined Category or Genre e.g. Comedy, Entertainment, Music, Travel or single words relating to VideoTag categories e.g. 'Glastonbury' and 'Festival' or 'Science'		
	B	Specific relationship between tag and video content		
<i>Iconographic</i>	Tag identifies what the video is of. Familiarity or some existing knowledge is needed to make this connection, sometimes an assumption may have to be made about this connection. Tags must not appear in textual data for the video e.g. the video title.		Specific	Objective
	1(a)	Tag specifically identifies what video is 'of' (place names/events) Tags which identify place names/events – a video of a concert tagged with the band name and venue, or a football match tagged with the team name, or an individual's holiday video tagged with the destination, requires knowledge acquired from familiarity with the specific place/event in question. Assumptions have to be made that a video tag is what it claims to be if the content is not familiar.		

	1(b)	Tag specifically identifies what video is 'of' (people/animals/objects) Tags which identify people/animals/objects – a video of Elvis Presley tagged as 'Elvis' or 'Presley' requires knowledge and familiarity of Elvis Presley; a video of a dog tagged as 'labradoodle' requires knowledge of the breed; or identifying specific objects such as 'fiesta' instead of 'car', 'hydrogen' instead of 'gas', 'hammer' instead of 'tool'. Distinctions cannot always be made between 'famous' people and 'non-famous' people, therefore the assumption has to be made that a video of a girl tagged as 'Sarah' is in-fact a video of a girl who is called 'Sarah'.		
	1(c)	Tag specifically identifies what video is 'of' (actions) Tags which identify specific actions of people, animals or objects in the video e.g. 'laughing', 'shouting', 'driving', 'hiding', 'kissing', 'barking', 'boiling', 'bouncing'. This category captures motion or behaviour that is not always identifiable in a still image.		
	1(d)	Multi-word Tags (phrases or whole sentences that describe the content additional to existing textual data) Multi-word tags contain two or more words, they must describe the content using words that do not appear in existing textual data for the video e.g. the title or refer to the owner of the video. If more than one word is used, even if the tag can be classified in another category it is classed as multi-word e.g. tags can include a person's full name 'Elvis Presley' an event 'Liverpool versus Everton' or phrases 'boat sailing on lake Windermere'.		
	Tags that interpret the video content at its most abstract and subjective level.			
<i>Iconological</i>	2	Tag identifies what the video is about Typically expressed by the use of abstract nouns or adjectives - an interpretation is made of what the video is about e.g., video of people smiling tagged as 'happiness'; video of cars on a racetrack tagged as 'speed' ; a video of a card trick tagged as 'skill'.		<i>Subjective</i>
	3	Tag which expresses opinion of the content Tags that denote the taggers opinion of the video e.g. 'funny', 'rubbish', 'boring', 'poo', 'cool'.		
	C	Tag only useful to a minority of users, specific individual or group		
	Tags which have a primarily social or categorising function.			
	1	Refining tag Tag which cannot stand alone - only useful when looked at as part of the larger tag set e.g., episodes of a series of videos specified by a number, acronyms or dates.		

	2	<p>Self-reference tag Tags which identify video content in terms of its relationship to either the tagger or the specific group which the video belongs to. Tags can be single or multi-word e.g., 'my dog'; 'our graduation'; 'holiday 2012' OR tags which appear useful, but show no relationship/connection to the accompanying video e.g. 'Sarah's song'.</p>	
	3	<p>Tag which explicitly denotes ownership of video Tags can be single or multi-word and refer to the owner of the video e.g., video tagged with the same username as that of the person who uploaded the video.</p>	
	D	Irrelevant/Non Useful Tags	
	Tags that are not useful as additional textual data describing the content for search or have no relevance to the video content.		
	1	<p>Compound tag Tags where words, phrases and sentences are joined together as one long text string e.g. 'dogchasingastick', 'davidcameron', 'westbromwich'</p>	
	2	<p>Multi-word Tags Tags containing two or more words that repeat textual data already available such as the video title or VideoTag category. All words in the tag must appear in the textual data otherwise use B1d.</p>	
	3	<p>Attention Attracting Tags Tags that are assigned to attract attention to the video, that refer to popular search terms, but have no relevance to the video content e.g. 'porn', 'sex', 'celebrity name', 'win', 'free'.</p>	
	4	<p>Misspelling Whilst it may be obvious what the tag is meant to be, a misspelling obviously renders the tag useless in terms of subsequent users of the system who are searching for videos with that specific tag, unless they too misspell the tag/word e.g. 'Belguim' instead of 'Belgium'.</p>	
	5	<p>Unable to determine relationship Despite having attempted to look up either the meaning of the tag and whether the tag is a foreign word or not, tags which do not fit into any of the above categories will be deemed as unable to classify (e.g., nonsensical words).</p>	
	6	<p>Foreign word/character Tags that are not in the English dictionary and are not identifiable as having relevance to the content; contain characters e.g. Chinese or Arabic or random characters e.g. '!\$%&'</p>	
	7	<p>Conjunctions and prepositions Parts of phrases that have been separated into single word tags e.g. 'the', 'in', 'of', 'and'</p>	

	8	Repeated tags Multiword tags containing the same word e.g. 'dog dog'	
	9	URL A link to a website e.g. http://www.youtube.com	

Category A and B tags describe content at an objective or a subjective level. Objective tags are divided by the specificity at which they describe objects in the video. Only two levels of basic level theory are modelled: basic and subordinate. Basic is the first level at which a person recognises an object (Rosch, 1975) (e.g., 'dog'). Tags describing content at this level will be in an A category. Users can expand their description to include the superordinate parent (e.g., 'animal') for this example, which will again be an A category, or the subordinate child (e.g., 'labradoodle') which will be a B category. It is likely in a tagging system that the tag 'dog' will be entered to describe the content, the tag 'animal' might be added as a general category and 'labradoodle' would only be added by a user with a specific knowledge of the breed. Whilst tags that relate to a category or genre could be classified as subjective as in Kim (2011), for the purpose of this research these tags are classified as basic A2. Tags that describe a category or genre do not create additional textual data, they reproduce the existing YouTube or VideoTag video category which is obvious to the VideoTag user. Subjective tags interpret the video content as a whole, and do not identify individual objects; B2 identifies what users perceive the video to be about, B3 captures user opinion of the video. Category C tags identify social communication behaviour. These tags are the least expected tag type due to a lack of community and social sharing in the VideoTag system. Category D tags classify instances of the vocabulary problem, malicious data and tags that bear no relevance to the video content.

7.4 Reliability of the Tag Classification Scheme

To test the reliability of the scheme a second classifier was used. 100 tags were selected at random from the VideoTag database, selecting from tags entered using each system during the testing phase and each experiment. The independent classifier was given detailed instructions to follow to help them classify the tags (see Appendix D).

7.5 Selecting Tags for Classification

All tags entered into VideoTag during the testing phases and two experiments were classified. The testing period has a different start date for each game: Golden Tag February 27th 2013, Top Tag March 13th 2013 and Simply Tag April 12th 2013. Each testing phase ends at the start of phase one (21st April 2013). Phase one continued until the start of phase two; phase two data is taken for the period 21st June 2013 – 22nd December 2013. Testing phase data encompasses the soft launch period before publicity; multiple users tried out the systems and entered valuable data. References to phase two in this chapter incorporate SciFest and phase two data because SciFest was a prototype of the phase two system as opposed to a unique experimental system; distinction is made by referring to phase two data with SciFest data removed. Four conditions were identified for comparing the tag type classification results:

1. Phase one compared to phase two – tag types entered into the two experiments were compared to discover whether design decisions made at the end of phase one affected the types of tag users entered in phase two. This investigates whether video content has an effect on tag type.

2. Golden Tag compared to Top Tag – tag types between the two games were compared to discover whether elements of gameplay had an effect on tag type.
3. Game compared to non-game – Tags entered into the two game systems Golden Tag and Top Tag were compared to the types of tag entered in the non-game system, Simply Tag to assess whether game elements affect tag type.
4. Entertainment videos compared to informative videos – Only tags entered during phase one can be evaluated in this condition. Videos were categorised in phase one based on the Entertainment and Informative category split used to classify YouTube and Viddler tags in Section 4.2. Tag types assigned to videos in each category, Entertainment and Informative were evaluated and compared to the findings reported in Section 4.2. The aim is to discover whether users exhibit similar tagging practises using VideoTag as for YouTube and Viddler and further investigates whether video content affects the tags users enter.

Table 7-2 shows the number of tags classified for each testing condition, a brief description of each condition and the amount of users who participated. In each case all available tags were classified. The analysis focuses primarily on how specifically users describe the video content. The amounts of basic and specific objective tags were compared between conditions as well as the quantities of subjective tags. Table 7-3 shows the tag type classifications categorised by objective or subjective groupings.

Table 7-2 The amount of tags classified in each condition.

Set	Tags	Users
Phase 1 <i>The first VideoTag experiment, Apr-June 2013.</i>	986	13
Phase 2 <i>The second VideoTag experiment, June-Dec 2013 inclusive of the SciFest prototype.</i>	3228	17 (plus many users using the SciFest and Guest accounts)
Phase 2 (no SciFest) <i>The second VideoTag experiment, June-Dec 2013 with data from the SciFest prototype removed.</i>	1018	17 (plus many users using the Guest account)
Entertainment <i>All tags entered during phase one for videos in entertainment categories (comedy, entertainment, music and gaming).</i>	1773	13
Informative <i>All tags entered during phase one for videos in informative categories (education, news, sport, technology and travel).</i>	905	13
Golden Tag <i>All tags entered using Golden Tag and Top Tag during the testing phase, phase one and phase two.</i>	5360	34 (plus many users using the SciFest and Guest accounts)
Golden Tag <i>All tags entered using Golden Tag during the testing phase, phase one and phase two.</i>	2723	30 (plus many users using the SciFest and Guest accounts)
Top Tag <i>All tags entered using Top Tag during the testing phase, phase one and phase two.</i>	2637	23 (plus many users using the SciFest and Guest accounts)
Simply Tag <i>All tags entered using Simply Tag during the testing phase, phase one and phase two.</i>	929	15 (plus many users using the SciFest and Guest accounts)

Table 7-3 Tag type categorised by Objective and Subjective Vocabulary

Vocabulary Category	tag type Category
Basic Objective	A1, A2
Specific Objective	B1a, B1b, B1c, B1d
Subjective	B2, B3

7.6 Inter-coder Test for Reliability

The level of agreement between the two independent blind classifications of 100 tags was calculated using a standard measure for inter-coder reliability, Cohen's kappa, giving .831. This indicates an excellent level of agreement (Landis and Koch, 1977; Fleiss et al., 2013). This validates the validity of the tags.

7.7 Results

7.7.1 General Observations

The VideoTag experiment generated 6289 tags during the testing periods and two experiments (phase one and phase two combined). The majority of tags (>80%) were relevant to the video and described its content. The most frequently entered tag types were A1 - basic description of what the video is of, B1b - specific description of what the video is of and B1d - multi-word tags that extend the existing textual data. There were more specific multi-word (B1b) tags than basic (D2) tags. In most cases a compound tag (D1) with multiple words joined together as a single word was also accompanied by the same word as a multi-word tag. This could be a typing error, or could be from a user with wide experience of tagging, as most tagging systems do not allow multi-word tags but allow compound tags. This could also be attributed to the hash tags used on systems like Twitter and Instagram. Very few tags had no

relationship to the video content (D5) and any malicious tags were in the form of random letters; these were classified as Misspellings (D4) rather than (D5) which was retained for tags that appeared relevant to the video but on watching the video could not be identified. Misspellings were the main aspect of the vocabulary problem identified in VideoTag tag data. Synonyms and plurals were not identified or classified as irrelevant because in terms of increasing textual data that describe the video content both are useful. There were no Conjunctions and Propositions (D7) in contrast to the YouTube data set (Section 4.1). This is because Multi-word tags were allowed as a single tag and not separated into individual tags. Videos in VideoTag are tagged only to describe the video content; the tagger has no personal relationship with the video. Self Reference tags that were present were assigned to Vlogs, where the owner of the video was evident from the content and referred to in the video itself. No tags were used to Attract Attention (D3), emphasising a lack of social communication or personal organisation motivations.

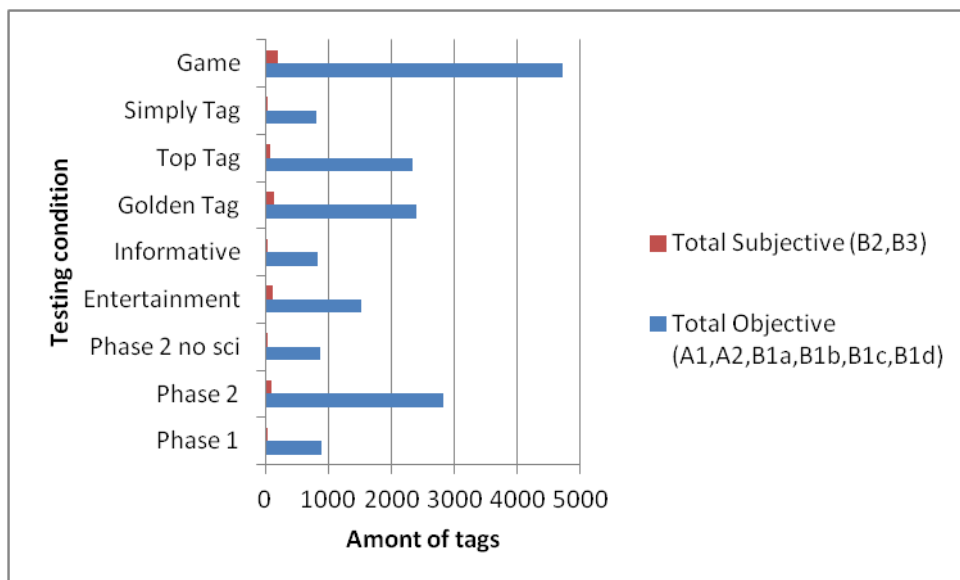


Figure 7-1 Overall distribution of subjective and objective tags

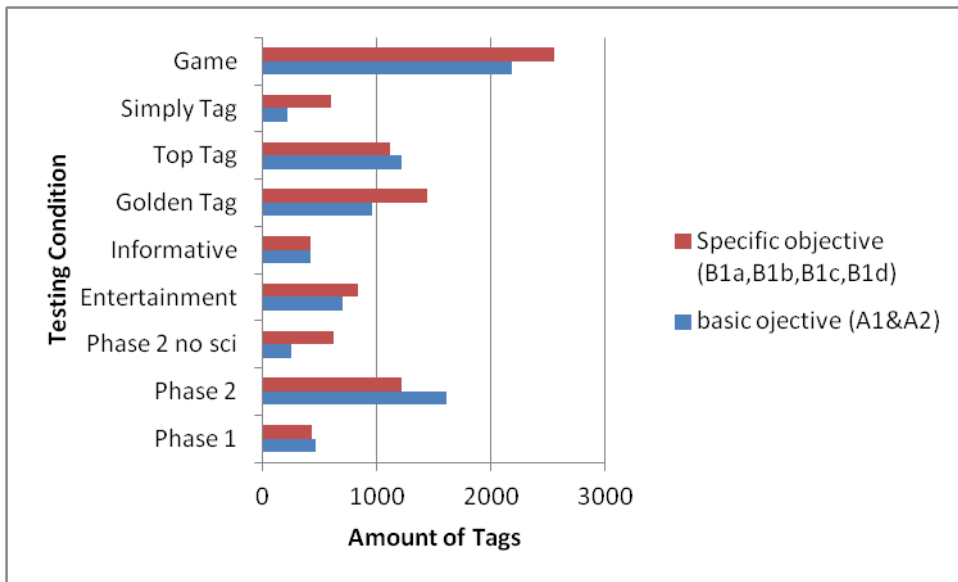


Figure 7-2 Overall distribution of specific objective and basic objective tags

More objective tags than subjective tags are in the datasets (Figure 7-1). A greater tendency for users to describe objects in the video at a specific rather than a basic level is clear shown in (Figure 7-2). This propensity to identify objects at varying levels of specificity rather than interpreting the video content will be discussed further in relation to each testing condition.

7.7.2 Phase One and Phase Two

The main influence that could create a variation in tag type between the two experiments is the video content. Phase one contained videos within generic YouTube categories and phase two contained videos aimed at special interest groups (see Chapter 5 for detailed discussion of the two experiments). In total 986 tags were entered during phase one and 3228 tags were entered during phase two, 2,210 of

these during the SciFest experiment. Users entered tags using Golden Tag, Top Tag or Simply Tag in both phases. During the SciFest experiment users could only tag videos in the ‘Crazy Science Experiments’ category and the majority of users were aged between 8 and 13. This may have had an effect on tag type. The proportions of tag type assigned to each experiment phase are detailed in Table 7-4.

Table 7-4 Percentages of tag types in the phase one and phase two experiments.

Tag Type	Phase 1	Phase 2	Phase 2 (no SciFest)
A1	42.9%	46.3%	19.1%
A2	3.7%	3.5%	7.5%
B1a	6.3%	1%	1.9%
B1b	22%	22.9%	44.1%
B1c	5.7%	5.5%	3.7%
B1d	9.3%	8.4%	15.5%
B2	0.7%	0.7%	1.3%
B3	2.4%	2.2%	2.5%
C1	0.5%	0.03%	0%
C2	0%	0%	0%
C3	0%	0.2%	0.4%
D1	0.3%	0.2%	0%
D2	1%	0.2%	0.3%
D3	0%	0%	0%
D4	4.8%	7.7%	2.7%
D5	0.3%	0.2%	0%
D6	0%	0%	0%
D7	0%	0.1%	0%
D8	0.4%	1%	0.5%
D9	0%	0%	0%

A higher percentage of basic ‘of’ A1 tags (46.3%) and misspellings D4 tags (7.7%) were entered in phase two, when data generated during the SciFest experiment with school children is included, compared to phase one (42.9% and 4.8% respectively)

and phase two with SciFest data removed (19.1% and 2.7% respectively). An abundance of spelling mistakes and basic language is probably a result of tagger age and not system design. For example, when comparing tag type distribution with the SciFest data removed (phase two (no SciFest)), then the most frequently entered tag type in phase two is specific 'of' B1b (44.1%), doubling the amount entered during phase one (22%). This increase in specificity of description can be explained by the change in content; users had more specific knowledge of the content that they were tagging. This change also created an increase in the amount of A2 tags, 3.7% in phase one and 3.5% in phase two compared to 7.5% in phase two with SciFest data removed. More users entered tags that described the category, but the categories were more specific to the content than the general categories used in phase one (e.g., 'Comedy' in phase one compared to 'Epic Fails' in phase two). Perhaps the category restriction for SciFest users to 'Crazy Science Experiments' had an effect on the amount of category defining tags they entered. The category titles contained more specific language; therefore A2 tags assigned to phase two videos have a higher specificity of language e.g. Glastonbury, Wimbledon and Download. These tags were categorised as A2 not B1a because they appear in the category and video titles therefore, they reinforce categories rather than describing in more detail what the video is of or about. As a result, more tags that specifically described place names and events (B1a) were recorded during phase one than phase two. In spite of this practice, few multi-word tags that reinforce the title or other textual data (D2) tags were recorded in phase two. Fewer B1d multi-word tags were entered during phase two (8.4%) than during phase one (9.3%) and particularly phase two (no SciFest) (15.5%) again highlighting the potential affect the age of SciFest users had on overall tag type and indicating that single word tags are easier to think of than multiple word descriptions. A similar proportion of Opinion Expression (B3) tags were entered during each experiment (see Table 7-4), however more tags that interpret

what the video is about (B2) were entered during phase two (no SciFest) (1.3% compared to 0.7% in phase one and phase 2). Overall, there were few instances of social communication, or the vocabulary problem across the two data sets with the majority of tags creating relevant descriptions of the video content.

More basic objective tags were entered during phase two (49.8%) than during phase one (46.6%) and this changed considerably when the SciFest data was removed (26.6%). With SciFest data removed, phase two had a larger proportion of specific objective tags (65.2%) than phase one (43%); this reduces to 37.8% with SciFest data included. More subjective language was recorded during phase one (3.1%) than phase two (2.9%). With SciFest data removed this trend is reversed with more subjective tags entered in phase two (no SciFest) (3.8%). Overall during both experiments users entered more objective than subjective tags with an increase in specificity during phase two (SciFest removed). (See Table 7-3 for Objective and Subjective groupings of tag type).

Table 7-5 Results of z-test for differences in proportion tests for each tag type category between phase one and phase two.

Tag Classification Category	Phase 1 (p1)	Phase 2 (p2)	Proportion p1	Proportion p2	z-test result (z)	z-test result (p)
A1	423	1494	42.9%	46.3%	-1.8665	0.061
A2	36	113	3.7%	3.5%	0.224	0.826
B1a	62	32	6.3%	1.0%	9.857	0.000
B1b	216	739	21.9%	22.9%	0.6478	0.516
B1c	56	178	5.7%	5.5%	0.1983	0.841
B1d	90	271	9.1%	8.4%	0.7193	0.472
B2	7	23	0.7%	0.7%	0.0084	0.992
B3	24	72	2.4%	2.2%	0.375	0.704
C1	5	1	0.5%	0.0%	3.4702	0.001
C2	0	0	n/a	n/a	n/a	n/a
C3	0	5	0.0%	0.2%	1.2366	0.215
D1	3	6	0.3%	0.2%	0.7048	0.484
D2	10	5	1.0%	0.2%	3.9653	0.000
D3	0	0	n/a	n/a	n/a	n/a
D4	47	248	4.8%	7.7%	-3.1409	0.000
D5	3	7	0.3%	0.2%	0.4957	0.624
D6	0	0	n/a	n/a	n/a	n/a
D7	0	3	0.0%	0.1%	-0.9576	0.337
D8	4	31	0.4%	1.0%	-1.6796	0.093
D9	0	0	n/a	n/a	n/a	n/a
Basic ojective (A1&A2)	459	1607	46.6%	49.8%	1.7765	0.075
Specific objective (B1a,B1b,B1c,B1d)	424	1220	43.0%	37.8%	2.9342	0.003
Total Subjective (B2,B3)	31	95	3.1%	2.9%	0.74896	0.324
Total	986	3228				

Key: grey rows show significant differences (p<0.05).

To assess the differences in proportions for tags of each tag type entered during phase one and phase two a series of difference in proportions tests were conducted for each of the 20 categories. For each tag type category the null hypothesis was that there is no difference in the proportion of tag type for tags entered during phase one and tags entered during phase two. Table 7-5 shows the results of a series of two-tailed z-tests using two independent data samples. A significant difference was found in the proportion of specific place name (B1a) tags with more being entered in phase one than phase two (p=.000). More refining (C1) tags were entered into phase one (p=.001). The video title was visible in phase two and as a result a larger proportion of multi-word tags that repeated existing textual data D2 were found in

phase two (p=.000). The increase in the amount of misspellings (D4) tags entered during phase two was found to be significant (p=.000). To evaluate differences in vocabulary between phase one and phase two tag sets, further two-proportion z-tests were conducted calculating differences in proportion of basic objective, specific objective and subjective tags entered during each experiment (see Table 7-5 for the tag type groupings). The null hypothesis in each of the three tests assumed that there was no difference in tag type between the two independent samples phase one and phase two. Table 7-5 shows the results; a significant difference in the proportion of specific objective tag types was found with a larger proportion entered in phase one than phase two (p=.003).

Table 7-6 Results of z-test for differences in proportion tests for each tag type category between phase one and phase two with SciFest data removed.

Tag Classification Category	Phase 1	Phase 2 (no SciFest)	Proportion p1	Proportion p2 nosci	z-test result (z)	z-test result (p)
A1	423	189	42.9%	18.6%	11.8248	0.000
A2	36	75	3.7%	7.4%	-3.6361	0.000
B1a	62	20	6.3%	2.0%	4.8843	0.000
B1b	216	441	21.9%	43.3%	-10.209	0.000
B1c	56	49	5.7%	4.8%	0.87	0.384
B1d	90	165	9.1%	16.2%	-4.7551	0.000
B2	7	12	0.7%	1.2%	-1.0828	0.280
B3	24	26	2.4%	2.6%	-0.1721	0.865
C1	5	0	0.5%	0.0%	2.2749	0.023
C2	0	0	n/a	n/a	n/a	n/a
C3	0	4	0.0%	0.4%	-1.9703	0.048
D1	3	0	0.3%	0.0%	1.7613	0.078
D2	10	4	1.0%	0.4%	1.6694	0.095
D3	0	0	n/a	n/a	n/a	n/a
D4	47	28	4.8%	2.8%	2.3774	0.017
D5	3	0	0.3%	0.0%	1.7613	0.078
D6	0	0	n/a	n/a	n/a	n/a
D7	0	0	n/a	n/a	n/a	n/a
D8	4	5	0.4%	0.5%	-2861	0.772
D9	0	0	n/a	n/a	n/a	n/a
Basic ojective (A1&A2)	459	264	46.6%	25.9%	9.6089	0.000
Specific objective (B1a,B1b,B1c,B1d)	424	675	43.0%	66.3%	-10.4804	0.000
Total Subjective (B2,B3)	31	38	3.1%	3.7%	-0.7227	0.472
Total	986	1018				

Key: grey rows show significant differences (p<0.05).

To investigate the extent to which SciFest data generated predominately by school children affected the types of tags entered during phase two the same procedure was repeated using the phase one data sample and the phase two data sample with SciFest data removed. The null hypothesis for each two-proportion z-test assumed that there would be no difference in the proportion of tag type for each of the 20 categories in each independent sample (Table 7-6). A larger proportion of basic descriptions (A1) tags ($p=.000$) and specific place names (B1a) tags ($p=.000$) were found in phase one. A larger proportion of basic category (A2) tags ($p=.000$); specific people/objects (B1b) tags ($p=.000$) and specific multi-word tags (B1d) ($p=.000$) were found in phase two, suggesting an increase in specific level tags in phase two with SciFest removed. A significant difference in proportion of tags which denote ownership (C3) were found with a larger proportion entered in phase two ($p=.048$). Increased visibility of textual data relating to the video in phase two lead to an increase in the number of A2 and C3 tags entered but interestingly no significant difference was found in the amount of D2 tags. This suggests that users in phase two (excluding SciFest participants) reinforced the video category and title using single word tags using multiword tags to further describe the video content and enhance existing textual data. With SciFest data removed there was a significantly higher proportion of misspellings in phase one in comparison to phase two ($p=.017$) indicating that SciFest users were responsible for the increase in misspellings in phase two.

Specificity of language was further investigated by conducting three further two-proportion z-tests, one calculating the difference in proportion of basic objective tag types and the second calculating the difference in proportion of specific objective tag types and third, the difference in proportion of subjective tag types. In each test the

null hypothesis assumed there would be no difference in proportions (Table 7-6). A larger proportion of basic objective tags were found in phase one ($p=.000$); In contrast, a larger proportion of specific objective tags were found in phase two with phase two with SciFest removed ($p=.000$). This suggests that general users of VideoTag, rather than SciFest participants, entered more specific level tags when more subject specific videos were included. Interestingly, there was no significant difference in the amount of subjective (B2 and B3) tags with users consistently entering few subjective tags across both experiments. Whilst content affects the specificity of language it does not encourage users to enter more subjective tags than objective tags.

Based on the results of these two proportion z-tests further tests of the phase two dataset was conducted to reveal differences between the proportion of basic and specific objective tags types. Table 7-7 shows the significant results from seven difference in proportion two-tailed z-tests. The B1c result is not shown as a significant difference was not found. More basic level tags were entered during SciFest (49.8%) compared to phase two (no SciFest) (25.9%), ($p=.000$). A larger proportion of A1 tags were entered during SciFest (46.3%) compared to phase two (no SciFest) (18.6%) ($p=.000$). More A2 tags were entered into phase two (no SciFest) (7.4%) compared to SciFest (3.5%) ($p=.000$). More specific objective tags were entered during phase two (no SciFest) (66.3%) than during SciFest (37.8%) ($p=.000$). Few B1a tags were entered in phase two. More B1b tags were entered into phase two (no SciFest) (43.3%) compared to 22.9% during SciFest ($p=.000$) and more B1d tags were entered during phase two (no SciFest) (16.2%) than SciFest (8.4%), ($p=.000$). The results clarify that more basic objective tags were entered during SciFest and that once the tags entered by SciFest users are removed phase two generated more

specific objective tags. This helps to quantify the finding that a change in video content to specific interest videos encouraged users to describe videos at a more specific level.

Table 7-7 Significant results from difference of proportion z-tests comparing basic and specific tag type classifications from phase two including SciFest data and phase two excluding SciFest data.

Tag Classification Category	Phase 2 (SciFest inc.)	Phase 2 (no SciFest)	Proportion p2 (SciFest inc.)	Proportion p2 (no SciFest)	z-test result (z)	z-test result (p)
A1	1494	189	46.3%	18.6%	15.7636	0.000
A2	113	75	3.5%	7.4%	-5.2293	0.000
B1a	32	20	1.0%	2.0%	-2.4619	0.014
B1b	739	441	22.9%	43.3%	-12.685	0.000
B1d	271	165	8.4%	16.2%	-7.1605	0.000
Basic ojective (A1&A2)	1607	264	49.8%	25.9%	13.3644	0.000
Specific objective (B1a,B1b,B1c,B1d)	1220	675	37.8%	66.3%	-15.956	0.000
Total	3228	1018				

Chapter 5 revealed that phase two generated less unique tags and yet tags were entered at higher specificity and at higher frequencies than in phase one. Golden Tag games in phase two generated less unique tags than did Top Tag, this is surprising as gameplay should create the reverse effect. The findings highlight that in phase two users began to agree more at a higher level of specificity, suggesting that the specific interest content encouraged users to tag at a more specific level. Interest in content improved specificity of tag type.

7.7.3 Golden Tag vs. Top Tag

Accounting for the slightly shorter testing phase for Top Tag, a similar amount of tags were generated by the two games (2,723 in Golden Tag and 2,637 in Top Tag) (Figure 7-3).

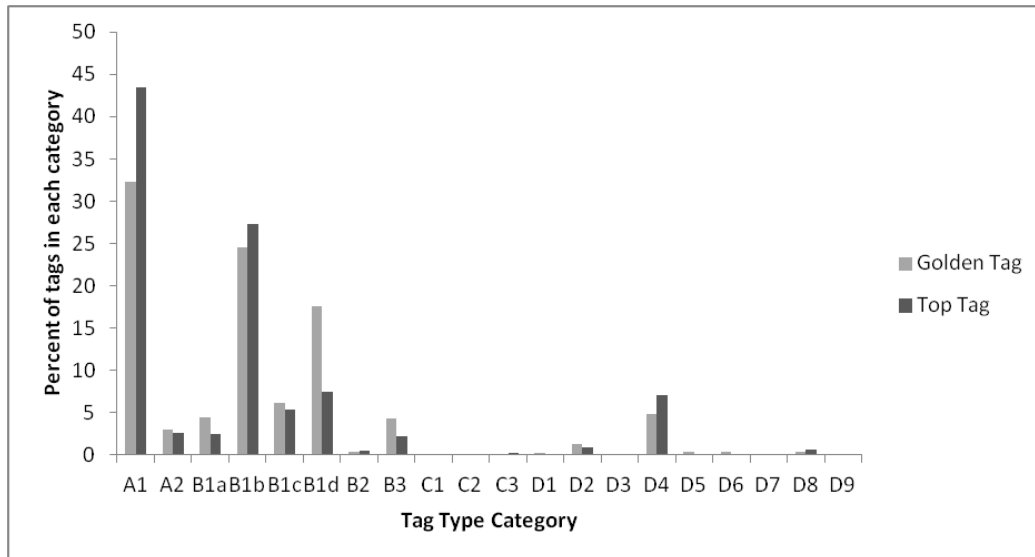


Figure 7-3 Distribution of tag type in Golden Tag and Top Tag

More basic 'of' (A1) and specific 'of' (B1b) tags were assigned to videos using Top Tag than Golden Tag. Conversely, in total more specific level tags were assigned to videos using Golden Tag with more specific place names and events (B1a), specific actions (B1c) and specific multi-word tags (B1d). More opinion expression (B3) tags were entered into Golden Tag although more subjective descriptions (B2) tags were assigned to videos using Top Tag. More non-descriptive (C and D) tags were entered into Top Tag and the largest proportion were misspellings, 7.1% in Top Tag and 4.8% in Golden Tag. This could be explained by the increase in use of Top Tag during SciFest. There was more repetition of tags (D8) in Top Tag where there is more

incentive to repeat a tag to see if it will score more points, although the amount was low compared to other tag types. In contrast, there were more compound (D1) tags entered into Golden Tag perhaps explained by users trying to find unique tags through form rather than language and more multi-word tags that reproduce textual data (D2). The overall increase in multi-word tags in Golden Tag (18.9%) compared to Top Tag (8.2%) reveals a tendency for users to favour single word tags in Top Tag, an indication that gameplay affects tag type as users are encouraged to enter tags that have high agreement. The vast majority of tags entered during both games described objects identified in the videos. In general, the proportion of objective language tags was similar in Top Tag and Golden Tag with more basic level language used in Top Tag and more specific level language used in Golden Tag.

To calculate the difference in proportion of tags of each tag type entered using Golden Tag and Top Tag a series of two-proportion z-tests were conducted for each of the 20 categories.

Table 7-8 Results of z-test for differences in proportion tests for each tag type category entered using Golden Tag and Top Tag.

Tag Classification Category	Golden Tag (GT)	Top Tag (TT)	Proportion GT	Proportion TT	z-test result (z)	z-test result (p)
A1	879	1146	32.3%	43.5%	-8.4384	0.000
A2	82	69	3.0%	2.6%	0.8733	0.384
B1a	119	64	4.4%	2.4%	3.9166	0.000
B1b	671	719	24.6%	27.3%	-2.1913	0.029
B1c	166	139	6.1%	5.3%	11.3036	0.194
B1d	479	194	17.6%	7.4%	11.3046	0.000
B2	10	13	0.4%	0.5%	-0.7041	0.484
B3	117	57	4.3%	2.2%	0.08914	0.000
C1	3	0	0.1%	0.0%	1.705	0.089
C2	0	2	0.0%	0.1%	-1.4374	0.150
C3	4	4	0.1%	0.2%	0.0454	0.960
D1	5	0	0.2%	0.0%	2.2015	0.028
D2	36	23	1.3%	0.9%	1.5781	0.114
D3	0	0	n/a	n/a	n/a	n/a
D4	131	188	4.8%	7.1%	-3.5868	0.000
D5	12	3	0.4%	0.1%	2.2651	0.023
D6	1	0	0.0%	0.0%	0.9842	0.327
D7	0	0	n/a	n/a	n/a	n/a
D8	8	16	0.3%	0.6%	-1.7157	0.085
D9	0	0	n/a	n/a	n/a	n/a
Basic objective (A1&A2)	961	1215	35.3%	46.1%	8.0369	0.000
Specific objective (B1a,B1b,B1c,B1d)	1435	1116	52.7%	42.3%	7.6061	0.000
Total Subjective (B2,B3)	127	70	4.7%	2.7%	3.9089	0.000
Total	2723	2637				

Key: grey rows show significant differences (p<0.05).

Table 7-8 shows the results of a series of two-tailed z-tests using the two independent data samples. A significant difference was found in the proportion of basic description (A1) tags with more being entered via Top Tag (43.5%) than Golden Tag (32.3%) (p=.000). A larger proportion of specific people/objects (B1b) tags were entered using Top Tag (27.3%) than Golden Tag (24.6%) (p=.029). A larger proportion of specific places/events (B1a) tags were entered using Golden Tag (4.4%) than Top Tag (2.4%) (p=.000) and also a larger proportion of specific multi-word (B1d) tags were entered using Golden Tag (17.6%) than Top Tag (7.4%) (p=.000). There was no significant difference between the proportions of general category (A2) tags, specific action (B1c) or subjective (B2) tags entered using either system. A significant difference in proportion was found in B3 opinion expression tags (p=.000), with a

larger proportion entered using Golden Tag (4.3%) than Top Tag (2.2%). Significant differences in the proportion of irrelevant tags were calculated, more compound (D1) tags ($p=.028$), more misspellings (D4) ($p=.000$) and more irrelevant (D5) tags ($p=.023$) were entered using Golden Tag (see Table 7-8 for the proportions). This suggests that encouraging users to enter tags that most other users have entered, like in Top Tag, encourages them to enter more relevant tags.

The design of Golden Tag encouraged users to enter more unique tags, with the design of Top Tag encouraging more tag agreement. The literature review highlighted that unique tags were of higher specificity, with users agreeing more on basic level tags. Therefore, Golden Tag should have generated more specific objective and subjective tags and Top Tag should have generated more basic objective tags. Comparing the groupings outlined in Table 7-3, Golden Tag generated more specific objective tags (52.7%) than did Top Tag (42.3%) and Top Tag generated more basic objective tags (46.1%) than did Golden Tag (35.3%). Correspondingly, Golden Tag had more subjective tags (4.7%) compared to Top Tag (2.7%). This suggests that the gameplay was sufficiently different to affect the types of tag users entered into each game and supports the view in the literature that users have higher levels of agreement with basic level tags. Further investigation of how gameplay affects the types of tag users enter was conducted by performing three two-tailed z-tests using the Golden Tag and Top Tag datasets as two independent samples. Difference in proportion of basic objective, specific objective and subjective tag types were calculated, see Table 7-8 for category groupings and results. The two-proportion z-tests found a larger proportion of basic objective tags were entered using Top Tag ($p=.000$). A significant difference in the proportion of specific objective tags was also found with a larger proportion entered using Golden Tag ($p=.000$). A significant

difference in proportion of subjective tags was also found, with more being entered using Golden Tag than Top Tag ($p=.000$). The results indicate that the types of tag users enter are affected by gameplay, with Golden Tag encouraging users to enter more specific level tags and subjective tags whereas Top Tag encourages users to enter more basic level tags.

7.7.4 Game vs. Non-Game (Simply Tag)

Considerably more tags were assigned to videos through the two game systems Golden Tag and Top Tag (5360) than in Simply Tag, the non-game system (929). In total more basic level category A tags were entered into games than Simply Tag. More basic 'of' tags (A1) were entered into the games (37.8%) compared to Simply Tag (18.7%). More category defining (A2) tags were entered into Simply Tag (4.6%) than into games (2.8%). In contrast, more specific level B tags were entered through Simply Tag as opposed to games even though more tags that specifically describe places and events (B1a) and opinion expression (B3) were entered into games. More specific 'of' (B1b), specific multi-words (B1d), specific actions (B1c) and specific about (B2) tags were entered into Simply Tag. There are few instances of the vocabulary problem (D) or social communication (C), the largest being misspellings (D4) in both game (6%) and non-game (7.1%). The misspellings in Simply Tag can be explained by an increase in malicious data in the form of random letters assigned to videos. Tags that looked authentic but were found to have no relationship to the video (D5) were not entered in Simply Tag and few examples were found in games (0.3%). Users had little incentive to corrupt the data, indicating that users were motivated by a desire to contribute to the VideoTag project rather than personal incentives to watch and share videos or engage in the games.

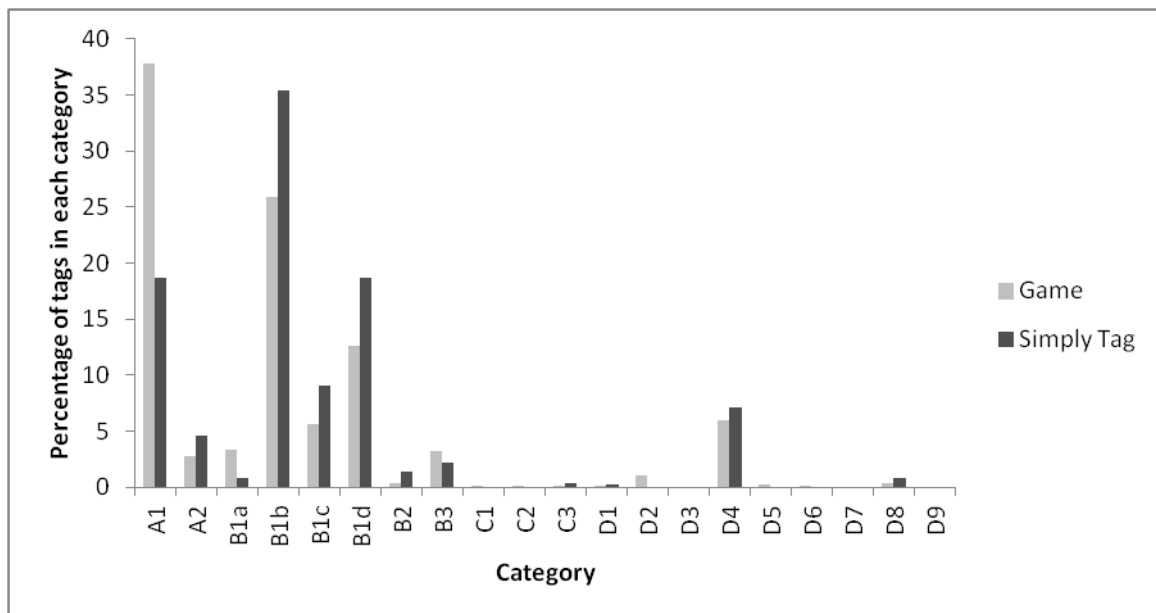


Figure 7-4 Distribution of tag type in Game (Golden Tag and Top Tag) and Non-Game (Simply Tag) systems

Differences in proportion tests were conducted using a series of two tailed z-tests using two independent samples, the proportions of tag type assigned using a game and the proportions of tag type entered into the non-game system, Simply Tag (Table 7-9). A significant difference in the proportions of basic level descriptions (A1) tags, ($p=.000$) and specific places/events (B1a) tags ($p=.000$) were found with a larger proportion being entered using a game. A larger proportion of basic category (A2) tags were entered using Simply Tag ($p=.003$). More specific level tags were entered using Simply Tag with significant differences in proportion of B1b ($p=.000$), B1c ($p=.000$), B1d ($p=.000$) and B2 ($p=.000$) tags. No significant difference was found in the proportion of opinion expression (B3) tags entered and no significant differences were found in the proportions of social (C) and irrelevant (D) tags.

Table 7-9 Results of z-test for differences in proportion tests for each tag type category entered using either a game or Simply Tag.

Tag Classification Category	Game (G)	Simply Tag (ST)	Proportion G	Proportion ST	z-test result (z)	z-test result (p)
A1	2025	174	37.8%	18.7%	11.2409	0.000
A2	151	43	2.8%	4.6%	-2.948	0.003
B1a	183	7	3.4%	0.8%	4.3739	0.000
B1b	1390	329	25.9%	35.4%	-5.9864	0.000
B1c	305	85	5.7%	9.1%	-4.036	0.000
B1d	673	174	12.6%	18.4%	-5.0988	0.000
B2	23	13	0.4%	1.4%	-3.6188	0.000
B3	174	20	3.2%	2.2%	1.7794	0.075
C1	3	0	0.1%	0.0%	0.7213	0.472
C2	2	0	0.0%	0.0%	0.5889	0.555
C3	8	4	0.1%	0.4%	-1.8139	0.070
D1	5	3	0.1%	0.3%	-1.8129	0.070
D2	59	5	1.1%	0.5%	1.3771	0.114
D3	0	0	n/a	n/a	n/a	n/a
D4	319	65	6.0%	7.0%	-1.2284	0.219
D5	15	0	0.3%	0.0%	1.6143	0.107
D6	1	0	0.0%	0.0%	0.4164	0.674
D7	0	0	n/a	n/a	n/a	n/a
D8	24	7	0.4%	0.8%	-1.2284	0.219
D9	0	0	n/a	n/a	n/a	n/a
Basic objective (A1&A2)	2176	217	40.6%	23.4%	9.9908	0.000
Specific objective (B1a,B1b,B1c,B1d)	2551	595	47.6%	64.0%	-9.2598	0.000
Total Subjective (B2,B3)	197	33	3.7%	3.6%	0.1846	0.857
Total	5360	929				

Key: grey rows show significant differences (p<0.05).

To further investigate the difference in proportion of basic objective, specific objective and subjective tags entered using either a game or Simply Tag two proportion z-tests were conducted for three cases (Table 7-9). More basic objective tags (40.6%) were entered into games than the non-game system, Simply Tag (23.4%), the difference in proportion was (p=.000). More specific objective tags were entered into Simply Tag (64%) compared to games (47.6%) the difference in proportion was (p=.000). The results indicate that users entered more basic level tags into a game environment and more specific level tags into a non-game environment when tagging videos as Goh *et al.* (2011) found with tagging images. This could be a result of time pressure causing users to enter more basic tags that are quicker to think of, as

Gligorov *et al.* (2011) suggested. There was no significant difference in the proportion of subjective tags entered, a similar amount were entered in both games (3.7%) and Simply Tag (3.6%). Even without the pressure of time few users entered tags that interpret the video content. The results suggest that applying game elements to a video tagging system affects how users tag videos. Users tag more specifically in a non-game environment than in a game environment.

7.7.5 Entertainment vs. Informative

Nine categories of video were available in phase one. Four were grouped as Entertainment (Comedy, Entertainment, Gaming and Music) and five as Informative (Education, News, Sport, Technology and Travel). Eight of the categories (excluding Education) mirrored the categories of video used for the tag evaluations in Chapter 4. Users preferred to tag videos that entertained them; overall more tags were assigned to entertainment videos (1773) than to informative videos (905) in VideoTag. The vast majority of tags were relevant to the content, providing descriptions that could enhance textual data for each video. To calculate the difference in proportion of tags of each tag type entered for entertainment and informative videos a series of two-proportion z-tests were conducted for each of the 20 tag type categories. Individual null hypotheses for each of the 20 tests assumed that there is no difference in the proportion of tag type entered for entertainment videos than for informative videos (Table 7-10).

Table 7-10 Results of z-test for differences in proportion tests for each tag type category entered for Entertainment and Informative videos.

Tag Classification Category	Entertainment	Informative	Proportion Ent	Proportion Inf	z-test result (z)	z-test result (p)
A1	626	393	35.3%	43.4%	-4.0928	0.000
A2	68	19	3.8%	2.1%	2.3967	0.016
B1a	39	87	2.2%	9.6%	-8.5701	0.000
B1b	431	174	24.3%	19.2%	2.975	0.003
B1c	135	45	7.6%	5.0%	2.5826	0.010
B1d	225	113	12.7%	12.5%	0.1505	0.881
B2	25	19	1.4%	2.1%	-1.3275	0.184
B3	89	19	5.0%	2.1%	3.6336	0.000
C1	3	1	0.2%	0.1%	0.3721	0.711
C2	0	0	n/a	n/a	n/a	n/a
C3	11	0	0.6%	0.0%	2.3744	0.018
D1	0	3	0.0%	0.3%	-2.4257	0.015
D2	35	0	2.0%	0.0%	4.2546	0.000
D3	0	0	n/a	n/a	n/a	n/a
D4	77	30	4.3%	3.3%	1.2848	0.201
D5	6	0	0.3%	0.0%	1.752	0.080
D6	0	0	n/a	n/a	n/a	n/a
D7	0	0	n/a	n/a	n/a	n/a
D8	3	2	0.2%	0.2%	-0.2937	0.772
D9	0	0	n/a	n/a	n/a	n/a
Basic ojective (A1&A2)	694	412	39.1%	45.5%	3.1728	0.002
Specific objective (B1a,B1b,B1c,B1d)	830	419	46.8%	46.3%	0.2527	0.803
Total Subjective (B2,B3)	114	38	6.4%	4.2%	2.3601	0.018
Total	1773	905				

Key: grey rows show significant differences (p<0.05)

More basic description (A1) tags (43.4% informative, 35.3% entertainment) were assigned to informative videos (p=.000). More specific places (B1a) tags were also assigned to informative videos (9.6%) compared to entertainment videos (2.2%) (p=.000). A larger proportion of basic category (A2) tags were assigned to entertainment videos (3.8%) than informative videos (2.1%) (p=.0164). There was a significant difference in the proportions of B1b tags (p=.003) with more assigned to entertainment videos (24.3%) than informative videos (19.2%) and B1c tags (p=.010) again with more assigned to entertainment videos (7.6%) than to informative videos (5%). Informative videos had a larger proportion of objective tags (91.8%) than did entertainment videos (85.9%) tagged at a more basic level. Informative videos had

similar amounts of basic objective (45.5%) and specific objective tags (46.2%), whereas more specific objective tags were assigned to entertainment videos (46.8%) compared to only 39.1% basic objective tags. There was a significant difference in the proportion of basic objective tags, with more being assigned to informative videos ($p=.002$), however there was no significant difference in the proportion of specific objective tags assigned to either category of video. More overall subjective tags were entered for entertainment videos (6.4%) than for informative videos (4.2%), a significant difference ($p=.018$) was recorded. However, users enter more opinion expression (B3) tags (5% entertainment, 2.1% informative) for entertainment videos and more tags that interpret the content (B2) to informative videos (1.4% entertainment, 2.1% informative). The same proportion of B2 and B3 tags were entered for informative videos. No significant difference was found between the proportion of B2 tags entered for each category of video, but a significant difference in the proportion of opinion expression (B3) tags was found ($p=.000$). A similar proportion of multiword tags that added additional textual data (B1d) were assigned to both entertainment and informative videos and although more multiword tags that reproduce existing textual data (D2) were added to entertainment videos, none were assigned to informative videos. A significant difference in the proportion of D2 tags was found ($p=.000$). More C3 tags that denote ownership were assigned to entertainment videos, a significant difference in proportion was calculated ($p=.018$). The title of the video and video owner were not visible to the users so any D2 or C3 tags are not a result of users entering what they see. This suggests that the titles of entertainment videos refer to the video using words that describe easily identifiable objects more than informative videos and that there are more instances of the video content referencing the video owner.

The results indicate that video content affects the types of tags users enter. The same findings reported in Section 4.2 when comparing the tag types entered for entertainment and informative videos on YouTube and Viddler were found in VideoTag tag data. Users entered more basic 'of' (A1) tags and specific place names and events (B1a) to describe informative videos than entertainment videos. More specific 'of' (B1b) tags and tags that describe specific actions (B1c) were assigned to entertainment videos than informative videos. Users entered tags of more subjective language for entertainment videos than informative. These results are consistent with tagging behaviour on YouTube and Viddler and validate VideoTag as a tagging system.

7.8 Discussion

Users enter more basic description A1 tags than any other tag category, which echoes the findings of Stuart (2012), followed by specific description B1b tags. Overall specific objective language is entered more than is basic objective language. Users describe objects in the video content at a specific level without expressing many opinions or interpreting what they are watching; there are few subjective tags compared to objective tags. Games generated more basic objective tags whereas Simply Tag generated more specific objective tags. Goh *et al.* (2011) suggested that games encourage users to tag for longer, tagging an image more than once, which in turn encourages them to enter more specific language tags. However, the temporal nature of videos means as the content changes consistently over the tagging period so users are not forced to think more specifically over time as they would be when tagging a still image. The results indicate that the time limit applied to games makes users enter more basic level tags which echoes the findings of Gligorov *et al.* (2011) although, users were encouraged to enter more specific level and subjective tags

using Golden Tag indicating that games can be designed to encourage more specific level and subjective language. Compared to the findings in Chapter 4, it seems that users tag videos with more specific terms than they do images. Stuart (2012), Hollink *et al.* (2004), Ransom and Rafferty (2011) and Jaimes and Chang (2000) all found that images were tagged with mostly basic level tags. Contrasting results were reported by Goh *et al.* (2011), who found that manual tagging of images created higher quality tags than did image tagging games although only the specificity of tags were classified with specificity measured by basic level theory; tags were not classified using the Panofsky/Shatford model. This research supports the findings of Goh *et al.* (2011); whilst users preferred to tag in a game environment, overall they entered more specific level tags in the non-game system. Despite being used less on average more tags per video were entered using Simply Tag (12) than either game (5). Whilst game elements encouraged users to tag, they did not encourage users to enter tags of higher specificity. Although in phase one, whilst the tags generated by Simply Tag offer more specificity of language they lack the quantity to create rich variety in descriptions. However, in phase two, although Simply Tag was used less, it generated more tags than either game which suggests that having an interest in content inspired users to enter more tags; a larger quantity of tags of varying levels of specificity were generated than in phase one. Stuart (2012) found that image content had more of an effect on tagging practice than user motivation, a view supported by Arends *et al.* (2012). The results of the phase two experiment also suggest that video content affected tagging practice on VideoTag more than did game elements. Stuart (2012) found no relationship between motivation to tag and the type of tags entered for images on Flickr. Few users were motivated to use VideoTag, but those who did entered a rich variety of tags. This research found that system design can affect tag type, in this study there were significant differences in

proportions of tag type entered in each game and non-game system and between the phase one and phase two experiments.

The quality of a tag is measured by how effectively it describes the content of the video to which it is assigned. As the majority of tags are objective one tag alone will only identify one object in the video. To assess the overall quality, the descriptive power of a collection of tags needs to be measured (Melenhorst and van Velsen, 2010; Chi and Mytkowicz, 2008; Goh et al., 2011; Gligorov et al., 2011). High quality tags are not necessarily specific with high level semantics. If all tags were high level and specific they would only describe the content to users with a specific interest and knowledge. A range of basic and subjective language is required for a useful collection (Gligorov et al., 2013). The wider the variety of tags and objects they identify (as well as the presence of a few subjective tags) the more likely a video will match a keyword search. To be useful for indexing the tag set needs to not be too general that it applies to all videos and not too specific that it applies to only a few (Rafferty and Hilderley, 2007). With greater specificity and also a range of basic and specific language the more chance users will enter search terms that match tags. VideoTag created a wide range of tags. Many basic level and specific level tags were created and a few subjective tags. The majority of tags entered in VideoTag were descriptive of the video content primarily classified as A or B tag type. Few social, self reference (C) tags were found in the VideoTag dataset and there were few instances of the vocabulary problem or other irrelevant (D) tags; the majority of the tags described the video content. This is in sharp contrast to the results of the preliminary investigation that found more social tags than descriptive tags in the YouTube dataset. By not having a social layer users were encouraged to tag primarily as describers and categorisers without any motivation to communicate

through tags or to organise content. The lack of social or personal tags that are not useful for search and a propensity of descriptive tags that provide additional textual data that could improve indexing is a positive outcome of the VideoTag project. Nevertheless, without social or personal motivation to tag, there was little motivation for users to use VideoTag, although the majority who did use VideoTag played many games. Game elements alone are not enough to motivate large numbers of users in place of social or personal incentives and without these motivations interest in content alone will also not encourage participation.

7.9 Conclusion

The two games available on VideoTag offer two different game experiences. Whilst both encourage the user to tag videos, Golden Tag was designed to encourage more specific level tags and Top Tag to encourage more basic level tags. The classification results extend the findings of Chapter 5, suggesting that Golden Tag and Top Tag generate different types of tag because of differences in gameplay. The high numbers of basic objective tags entered into Top Tag, coupled with the finding from Chapter 5 that tags entered into Top Tag had higher agreement rates, supports the view in literature that users have higher levels of agreement with basic level tags. Even though few examples of irrelevant tag types were recorded in either game, fewer irrelevant tags were entered into Top Tag. Mimicking the matching strategy employed in the ESP Game model by encouraging users to enter tags that most other users have agreed upon is likely to create more tags relevant to the video content. Encouraging users to think of obscure and highly specific tags in Golden Tag created more instances of misspellings and social tags. The two games were used a similar amount of times, therefore as both encouraged different types of tag both games are useful at encouraging users to create a range of descriptive vocabulary for video as

opposed to using only one game. The results show that gameplay can be designed to encourage users to enter more of a certain tag type.

Considering the tag output from both Golden Tag and Top Tag combined, games generated more basic objective tags whereas the non-game system Simply Tag generated more specific objective language. Objects were more likely to be described at a basic level through games and at a specific level using Simply Tag. A lesser amount of subjective language was used in all systems. Few tags interpreted what the video was about at an abstract level (B2) and users rarely used tags to express opinions about the videos they watched (B3). A similar proportion of subjective tags were entered into the games combined and Simply Tag with more B2 tags entered using Simply Tag. More evidence of categorising behaviour was found in Simply Tag with an increase in basic Category A2 tags. However, the proportion was still low at only 4.6%. In both systems users were more likely to describe objects in the video than categorise or interpret the whole video content. Both types of system generated similar amounts of descriptive tags (A and B) compared to social or irrelevant tags (C and D); few instances of malicious or irrelevant data were recorded. More misspellings were expected in game tag data due to the time restraint, but, whilst a significant difference in proportion was found, more misspellings were entered using Simply Tag. Simply Tag was more open to malicious data and some of the misspellings can be explained as random letters being assigned to videos. It seems that the game environment and mainly the scoring system, actively discourages users from entering malicious tags, as suggested by Von Ahn (2006). Users may tag more specifically in a non-game environment but users prefer to tag in a game environment and so games generate more tags. Offering users a choice of system helps to generate a wider range of tag type. The fact that tag type varies between

systems indicates that using one tagging system limits the range of tag types. Video tagging games can complement an existing non-game video tagging system.

For the VideoTag experiment a video tagging system was created that used YouTube videos. The tagging games were not applied to an existing system with an active user base. Chapters 5 and 6 reported how this affected user motivation and limited levels of participation. The phase one experiment used videos from generic YouTube categories and focussed primarily on the tagging games. The phase two experiment created specific interest categories of video and changed the focus of the website from the tagging systems to the video content. Differences in tag type were evaluated through tag classification of the two datasets. Comparisons of phase one and phase two tag classification results did not reveal many significant differences in tag type proportion. Users appeared to use very similar language to tag videos in each experiment, with a tendency to use specific objective tags more in phase one. Chapter 5 reported that users of the SciFest prototype generated tags of high agreement. As high agreement tags are usually at the basic level, the phase two dataset with SciFest data removed was also classified. SciFest had a considerable effect on the overall tag types generated by phase two. The predominant effect was an increase in basic description (A1) tags and misspellings (D4) and when SciFest data was included more basic objective language was present. With SciFest data excluded phase two generated an increased amount of specific level tags compared to phase one with the majority being at specific objective level. Few subjective tags were generated, with similar proportions reported in both experiments. The findings of Chapter 5 indicated that a change in system design to emphasise content over games affected the types of tag users enter and this was confirmed by the tag classification. Emphasising use of VideoTag as a portal to curate specific interest videos (phase

two) rather than a portal for playing video tagging games (phase one) changed the specificity of the tag types users entered, with a decrease in the amount of A1s and an increase in the amount of B1bs and B1ds. The structure of the phase two website created category titles from the specific themes in the video content (e.g., 'Epic Fails', 'Greatest Wimbledon Moments', 'Hitler finds out...') this encouraged users to produce more basic category defining tags than when generic YouTube categories were used but few were entered. Few subjective tags were entered explaining that users prefer to identify objects to form descriptions rather than categorise and interpret the video content.

Whilst specific interest content affected tag type in phase two, generic categories also affected the types of tag entered in phase one. Videos were assigned to one of four entertainment categories or one of five informative categories. Users preferred to tag videos that entertained them in a game environment. There was no difference in the amount of specific objective language assigned to videos in each category however, more basic objective language was assigned to informative videos and more subjective language assigned to entertainment videos. Informative videos generated more A1 and B1a tags, entertainment videos generated more A2, B1b and B1c tags. This distribution of tag type was also reported for tags entered into YouTube and Viddler for the same video category split. The classification study found that the content of the video had an effect on the types of tag users enter. Users enter tags at a more specific level for entertainment videos and they are more likely to interpret the whole video or express opinion through tags if the video entertains them. Users enter more specific level tags if they have a specific interest in the content, but specific interest content does not encourage more subjective language.

In Simply Tag and phase two where the title of the video is visible still few basic multiword tags (D2) were entered, indicating that users were motivated to tag videos to create additional textual data rather than reproducing it. The high amounts of multi-word tags indicate VideoTag users were tagging as describers rather than as categorisers. This is enforced by the lack of A2 tags and is supported by the results in Chapter 5 that few tags had high frequency; there was little agreement on terms. It is a positive finding that VideoTag encourages users to be describers as the aim was to encourage users to create tags for additional textual data. Tags with high agreement can be used as categories for browsing and the whole tag set can be used to create semantically rich descriptions for videos. If a system uses tags to categorise and browse videos then multi-word tags are unwarranted. If tags are used to create descriptions to enhance textual data for search then multi-word tags are useful as they create natural language phrases that can be indexed and matched in keyword search or read by screen readers to improve accessibility (Von Ahn et al., 2006). It is a positive result that VideoTag has generated many multi-word tags.

Overall most users tag video in VideoTag with objective language that is relevant to the video content describing what they see. For each testing condition over 80% of the tags were objective at varying specificity, with more specific objective than basic tags. Objective language is better for search as more users will agree on basic terms. Vallet *et al.* (2008) and Kofler *et al.* (2012) found that the majority of users use basic level terms. This is because users simplify videos searches to achieve success because the textual data available for indexing is limited (Halvey and Jose, 2012; Tjondronegoro et al., 2009). However, de Rooij *et al.* (2008) note that users struggle to simplify their search queries enough. Therefore having textual data that contains more specific objective vocabulary would enable search for specific interest video. It

was difficult retrieving videos for the specific interest categories used in phase two using the YouTube API, the more specific the category the less videos were retrieved and queries had to be amended to a broader topic. This problem was also recorded by Marchionini *et al.* (2009), Paolillo and Penumathy (2007) and Tao et al. (2012) who found searching for special interest content difficult; by increasing specificity of language in the textual data visibility of these videos will be improved. Tags generated using VideoTag could be used to enhance the textual data for videos in YouTube or other video libraries thereby improving their visibility in search results.

8 Conclusion

8.1 Introduction

The primary aim of this research was to investigate whether game elements that make casual games engaging could be applied to a video tagging system to encourage users to tag online videos with tags that accurately described the video content. The rise in popularity of smartphones and tablets and increased mobile and wireless internet speeds has led to an increase in the amount of user generated video that is uploaded and the amount of online video that is consumed. The majority of this video is poorly labelled and poorly described however, making videos hard to find. By improving the textual descriptions assigned to videos using tags, it is possible that text-based video search could be improved, particularly for special interest or niche content. At present, no video sharing system exists that uses social tagging. VideoTag was created as an experiment to investigate whether users could be encouraged to tag videos through the use of video tagging games and whether they would enter tags that described the video content using a range of basic, specific and subjective vocabulary. This chapter summarises the key findings of the research, discusses the contribution to knowledge and outlines areas for future work.

8.2 Research Objectives

8.2.1 General Objectives

The general or background research questions were to establish why people play video tagging games, why they tagged, how people search for video, how they use YouTube and how they tag on YouTube. These questions were mainly answered through the literature review. Gaps in knowledge identified were: a lack of knowledge about how users tag videos on video sharing websites, how users tag videos using video tagging games and how game design theory can be applied to the design of video tagging games. Few video tagging games existed at the start of this research, VideoTag version one being the first video tagging game (Greenaway, 2007), followed by the Yahoo video tagging game (Zwol et al., 2008). There are currently few published findings about video tagging games. Some research has been conducted in parallel with this thesis but focusing on fragments of video, rather than the whole video, from specific professionally generated content (Dutch TV archives) and from the angle of comparisons with professional indexing rather than improving user tags. Game elements have not been researched, however.

The aim of this research was to add to the small body of research on video tagging games by investigating new methods to create the games that did not follow the ESP Game model, using elements of casual game design. The emphasis was on the tags that users create, with games designed to alter tagging behaviour. In order to assess the tag output generated by VideoTag an understanding of how users tag videos in video sharing sites was required. Preliminary studies were conducted to address this issue. Two video sharing sites were chosen: YouTube, the most popular, and Viddler.

In YouTube only the user uploading a video can tag it, but Viddler allows all users to tag a video. Tags on each site have different functions. Viddler videos are only organised by the tags that users enter, and so the tags categorise content more than they provide extra textual descriptions. In contrast, tags only provide additional textual data in YouTube. Popular videos on YouTube are likely to contain more tags (Halvey and Keane, 2007) but the preliminary studies found only a weak relationship between the number of views and the number of tags in both YouTube and Viddler. This supports Halvey and Keane's (2007) theory that popularity of a video is dependent on internal promotion metrics more than on search functions. If the textual data describing videos was improved by tagging then this would improve the performance of searches and video popularity may become more dependent on what users want to watch rather than what the system recommends. This thesis found no evidence of collaborative tagging on Viddler, despite it being possible in the system. Viddler videos seem to be mainly tagged by owners in an attempt to maximise the video audience, despite tags having little effect on the number of views that a video receives. Tagging is also not used to create personal collections of videos. Users tag Viddler videos with slightly more descriptive language and YouTube videos with more social tags. Using tags as the sole means of categorisation in Viddler apparently motivated users to enter relevant tags. Poor tagging behaviour was observed in both systems however, highlighting the need to improve the quality of tags for videos.

8.2.2 Specific Objectives

Four specific research questions drove the main part of the research. The findings for each question and how each research problem was resolved are discussed in this section.

8.2.2.1 R1 - Which game elements that make casual games engaging also help to make video tagging games engaging?

The primary goal of this research was to design video tagging games as a type of GWAP, using elements of casual game design theory to help design engaging games with the purpose of descriptively tagging videos. During the research, gamification began to have an academic impact, transitioning from an industry buzzword to a defined academic concept. It raised the question of whether VideoTag was a game or a gamified video tagging system. This distinction is important because users are motivated to use each type of system differently. A game is a standalone system that is played for the sole purpose of playing a game. If the game elements are removed then there is no incentive for it to be used. Game players are intrinsically motivated, having different personal goals that they want to accomplish by playing. In contrast, a gamified system is used for a specific task or outcome that is not reliant on the completion of certain game conditions or to follow game rules. Game elements such as collecting achievement badges, point scoring and leaderboards are there to promote competition between users but tasks can be completed without interacting with them. If the game elements are removed then there is still an incentive to use it, for its outcome, but interactions are presumably less enjoyable. Gamification may start with an element of intrinsic motivation, in the sense that some users may feel motivated to play based on the outputs produced, and then game elements can be incorporated in order to motivate users to use the system and to use it effectively. Nevertheless, the results suggested that the addition of extrinsic motivations, such as a scoring system, can deter users from using a system that they are intrinsically motivated to use. A gamified system may only attract people who are interested in the purpose or main activity of the system.

There is no definitive list of game elements for gamification. There are also no specific features that something must contain to be classed as a game. The elements must be chosen to promote the completion of certain goals without a set formula (Deterding et al., 2011b). If game elements are integrated well then it should not be possible to determine whether a system is gamified or is a game without knowing the designer's intention (Deterding et al., 2011a). Gamification in its simplest form is the application of a scoring system into an existing system but more creativity and a variety of game elements are required to properly incentivise users, especially in the current climate with a glut of mobile games. For example, adopting the ESP Game model for new tagging games would create a re-versioning of a system that some people have used before and which was successful in a less competitive environment.

Initially, Golden Tag and Top Tag were designed to be pure games in the sense of not considering the need to foster intrinsic motivations to play them. In phase two, the games were redesigned and repackaged as tools to tag special interest content, and so the focus of the site changed from game playing to content creation. Golden Tag and Top Tag both have a scoring system but also include challenges, a time limit, 'juicy' feedback, cartoon fiction, clear goals and obstacles. The constraints of the need for a tagging system were barriers to the game design and hindered the amount of game elements and playful interactions that could be designed. It is difficult to create playful interactions when they are restricted by the tag entry procedure, which has a significant cognitive load. Moreover, playful interaction cannot undermine usability, because it is important to casual game players. They will not work around a usability problem and so the game difficulty must only come from gameplay. The VideoTag interface was made more playful through the graphics, the fiction and feedback.

The individual games were designed with play in mind. A play-centred approach was used, based upon player types and casual game design theory and applying elements of these to the basic interface of a video player and a tag entry form. A combination of play centric and user centric approaches were used. Levels and in-game challenges were created for Achievers, Explorers might enjoy perusing the video library using the tags and Collectors could collect points, tags and collections of videos. However, there were not enough game elements to keep these player types engaged. Users looking for Hard Fun are unlikely to enjoy VideoTag as there is insufficient challenge, strategy and reward. Easy Fun is only provided by browsing tags and videos for users who find enjoyment in exploring the system to see if they can contribute to it, break it or modify it. VideoTag provided Serious Fun for people happy to find enjoyment in a useful task. Without a social layer, VideoTag offered no People Fun which can benefit games lacking in other fun keys (Lazzaro, 2004). Play could come from the VideoTag site as well as from the two games, but this would only appeal to a limited audience (Lin et al., 2008).

Clear goals were created for each game in addition to the overall purpose of the system. The goal of Golden Tag was to hunt for any tag that only one other user had entered. The goal for Top Tag was to find the five most frequently entered tags. The playability questionnaire revealed that users could tell the difference between the goals of both games. Users were motivated by the purpose and the challenge. Users preferred the easier challenge of finding popular tags (Top Tag) than finding a needle in a haystack (Golden Tag). Users felt motivated by the time limit, which added an additional element of challenge to the games. Site-wide levels, as well as levels within the individual games also created challenge. This also catered for player curiosity; in both games reaching a new level was rewarded by new video content

and in Golden Tag, a change in the fiction. Points scored in individual games were aggregated and users placed on a level thermometer, giving the challenge to reach the top and climb through the ranks of a fictional TV company. As a reward for reaching higher-level ranks, users unlocked the ability to upload videos to VideoTag. This gave them more control over the system, allowing them to tag videos they wanted to watch and make use of the tags entered for their own uploaded videos. No users interacted with this service, however. Without a community of users or a social layer, there was less motivation to reach the top. It could also be due to the static nature of the system; there is no dynamic content because of a lack of community and social layer. Gamification will only work if users are introduced to a meaningful community with similar interests (Groh, 2012). Whilst this was provided in a simplistic form, users did not engage with it. User studies showed that whilst users felt that their actions had an impact on their score, most were fairly neutral about the level of challenge and how the challenge progressed over time. Users showed little interest in the leaderboard because no users progressed to the high levels that unlocked new functionality in the website. The developer evaluation highlighted challenge as a problem; there was no increase in challenge as users progressed through the levels. Some users felt bored whilst playing VideoTag and few users said they would continue to play. This is understandable if the level of challenge does not increase throughout the game because the balance between boredom and anxiety is not well balanced once players' skills improve. Most users felt that VideoTag catered for users of different skill levels, but users never felt intellectually stimulated by VideoTag. Hamari *et al.* (2014) suggest that users might not spend enough time within a service in order to become interested in it. The overriding problem for video tagging games, as for GWAP in general, is that they are not games. A user is unlikely to select a GWAP if they are looking for a casual game to fill time or to relax after work. People will select a GWAP if they have an altruistic interest in the purpose, or

are motivated by the novelty to 'just try it out'. This is why the majority of GWAPs are played by many users only once (Von Ahn and Law, 2009). This was not true for VideoTag, however. Despite low user numbers, most users played many games. The fact that most users played more than one game indicates that some playful behaviour emerged. This suggests that the game elements employed did engage players, but only for a short time because there was little long term use. This is potentially because of the lack of a social layer and a lack of perceived use for the tags.

Video tagging game users are more likely to be motivated by the purpose of tagging videos than by playing a game. User studies have revealed that GWAP users rated usability higher than enjoyment and VideoTag scored higher for usefulness than for appeal or absorption. It was clear that users perceived VideoTag as a GWAP rather than a video tagging system or a casual game. Users understood that the purpose of VideoTag was to enter keywords that described the video and felt encouraged to do this. Most VideoTag users were motivated by altruism (to help with the specific project), but this motivation will attract few users in the long term. To test for gamification of the system deterring users interested in the tagging process itself Simply Tag was developed. This was the least used system, and so users clearly preferred to tag videos using games. Thus the game elements encouraged use and the gamification did not deter users. Low numbers of participants suggests that the purpose was not clear or not of interest to many people, or that not enough people heard about VideoTag, despite the publicity for it. A lack of perceived use for the system had a big impact on participation. VideoTag used YouTube videos, yet the tags were not fed into YouTube and so users could see no benefits from the tags that they had entered. Some potential players did not think that they had the skills to

complete the challenges and did not find the goals to be clear enough. Users had to read instructions before they could play and the purpose of the system had to be explained. This is a deterrent for creating the critical mass of players that is essential for a community to form. The results suggest that video tagging games have a limited appeal, but that users with an interest in the medium will engage with them to some extent, although perhaps not in the long term.

The goals of VideoTag were to attract game players and to engage them in a useful task, and to attract taggers and make the process of tagging more entertaining for them. This would create a community and allow its members to compete against each other. This was a problem because players and taggers will have very different motivations to participate and perhaps the methods used to appeal to taggers deterred the game players and vice versa. Whilst game players are easy to promote to, the market is saturated with games. Tagging is a system-specific activity to aid users to ultimately get more out of the system. This motivation is absent from VideoTag because users cannot make use of the tags they create for YouTube videos in VideoTag on YouTube itself. Although based upon a small sample of users, this research suggests that it was not the quality of the system that deterred its use. Instead, the visibility, perceived use and perceived sociability of the system was not enough to attract enough users.

8.2.2.2 R2 - Can game elements affect the types of tag that players enter?

The VideoTag project developed two tagging games, Golden Tag and Top Tag, and one non-game based tagging system, Simply Tag. As discussed above, users preferred to tag in a game environment than in a non-game environment. Games

were used more than Simply Tag and generated more tags. The absence of social sharing and the emphasis on the purpose of the tagging encouraged users to tag more descriptively. Overall, tags entered into VideoTag contained more objective than subjective language, were descriptive rather than social, and the majority were relevant to the content. In contrast, the preliminary studies found that YouTube tags were mostly social or irrelevant (C and D). Although Viddler tags were more descriptive than were YouTube tags, there were still more social and irrelevant tags than were generated by VideoTag. Social tags tend to be more useful for the individual that assigned the tag or small specific groups of users rather than being useful as additional textual data that describes the video content. It is therefore a positive finding that VideoTag generated few social or irrelevant tags.

VideoTag tags differed in how specifically they described the video content, with more basic objective tags entered using the games and more specific objective tags entered using Simply Tag; significant differences in proportions were found for both ($p=.000$). Users preferred to identify objects in the video rather than to interpret the content and create subjective tags, which is more difficult. Similar proportions of opinion expression tags (B3) were entered using both systems, but users of Simply Tag were more likely to interpret the video content and enter tags that described what the video was about, with a significant difference in the proportion of B2 tags ($p=.000$). Games generated more tags that described what the video was of at a basic level (A1) and more tags that described specific places and events (B1a, $p=.000$). In contrast, Simply Tag generated more basic category tags (A2), more tags that specifically identify people, animals or objects (B1b), more specific actions (B1c) and more multi-word tags that add to the existing textual data (B1d). The differences in proportion were all significant ($p=.000$ ($p=.003$ for A2 tags)). More misspellings (D4)

were entered using Simply Tag, which suggests that game elements can help to reduce the amount of malicious or irrelevant tags entered into the system, as suggested by Von Ahn (2006).

Although Simply Tag generated more specific level tags, far more tags were entered into the games and despite the above, games generated descriptions with a wide range of degrees of specificity. Gameplay was designed so that the two games should encourage different levels of vocabulary. Golden Tag was expected to generate more specific level tags and Top Tag more basic level tags. Tags that many users enter are more likely to be basic level. In Top Tag users were encouraged to enter tags that they thought would have a high level of agreement, but in Golden Tag they were encouraged to find unique tags. The design was successful at creating individual games that generated tags at different specificity levels. Top Tag produced more basic objective level tags ($p=.000$) with users entering more tags that described what the video is of at a basic level (A1) ($p=.000$). Top Tag users concentrated on entering tags that they thought other users would have entered, and the most frequently entered tag types overall were A1 (basic level description) and the B1b (people, animals or objects). Top Tag generated more B1b tags than did Golden Tag ($p=.029$), suggesting that VideoTag users agreed at a more specific level rather than just on basic tags. Golden Tag produced more specific level and more subjective tags than did Top Tag. There was a significant difference in proportion between specific objective and subjective tags ($p=.000$) entered in to the two games. Users of Golden Tag entered more B1a, B1b and B3 tags ($p=.000$).

Although both games used the same game engine and followed the same fundamental procedure (watch a video, enter tags that describe it and score some points before the timer runs out), the fiction, juiciness and goals of both games were different. The clear differences in tag type between the two games, follow their respective design goals, which shows that game elements have an effect on tag type. Thus it is clear that game elements *can* be used to encourage users to enter more tags of a certain type. This is useful if the games are to be used to create tags that compliment descriptions generated by either automatic methods or professional indexers, when basic level tags are not required; or to compliment user generated tags on social sharing sites like YouTube where more descriptive tag types are required rather than social tags.

8.2.2.3 R3 - Does video content affect tag type?

In the preliminary studies videos were categorized as being either entertainment or informative. The types of tag assigned to videos in each category were compared. The results indicated that the category of video had an effect on tag type. More specific level tags were assigned to entertainment videos and informative videos were mainly tagged at the basic level despite also being assigned many specific level tags. More tags that identify objects at a basic level (A1) and categorise at a basic level (A2) were assigned to informative videos. Nevertheless, users also assigned more tags that specifically described places and events (B1a) and tags that interpreted what the whole video was about (B2). More tags that describe people, animals and objects (B1b) and more opinion expression tags (B3) were assigned to entertainment videos as well as more social tags (C).

The same category split was used during phase one of the VideoTag experiment. In VideoTag, similar proportions of specific objective language were entered for both category of video, with more basic objective tags being entered for informative videos and more subjective tags being entered for entertainment videos. With the exception of generic category defining (A2) tags, which could be a result of differences in system design, users tagged entertainment and informative videos with the same types of tag using Video Tag as they did using YouTube and Viddler. More A1, B1a and B2 tags were assigned to informative videos and more B1b, B3 and social tags were assigned to entertainment videos. This provides a clear indication that the content of a video has an effect on the types of tag that users enter. This was investigated further during the phase two experiment.

The phase two redesign concentrated on video content rather than individual games. Games were not the focus of the website but were a tool to tag video content for a specific interest. This change of focus meant that more tags per user were entered into Simply Tag than using a game although games were still the preferred tagging method. An interest in content was a stronger motivation to tag than were game elements, but not necessarily for using the system as the number of participants was still low. More specific language was used to tag specific interest content as opposed to generic content, but there was no difference in the number of subjective tags used. More basic objective tags were entered for videos in generic categories (phase one) and more specific objective tags entered for videos in specific interest categories (phase two excluding data from the SciFest experiment). Generic videos in phase one were assigned more A1 tags that describe objects in the video at a basic level and more B1a tags that describe places and events at a specific level. In comparison, for specific interest videos in phase two users assigned more basic category A2 tags,

more B1b tags that describe people, animals and objects at a specific level and more multi-word tags that add descriptions to existing textual data (B1d). The amount of B1b tags assigned to videos more than doubled for specific interest content compared to videos in generic categories. When a user has a specific interest in the video content they are apparently compelled to describe objects in the videos at a more specific level. Category B tag types require the user to apply existing knowledge to identify the objects; being more knowledgeable about the video content meant that users entered more specific level tags. Nevertheless, an increase in familiarity with the content did not encourage users to interpret the content and use more subjective language.

Stuart (2012) found that the types of tag that users entered were affected by image content rather than by motivation to tag. The idea that content has an effect on the tags users enter has also been suggested by Goh *et al.* (2011), Gligorov (2012) and Arends *et al.* (2012). Hildebrand *et al.* (2013) suggest that users are more inclined to tag popular videos. This research supports these findings in the sense that through three different studies in this thesis there is a clear indication that users entered different types of tag depending on the video content. More tags were assigned to entertainment videos than to informative videos in YouTube, Viddler and VideoTag and more tags were assigned to specific interest videos using Simply Tag than using games. User numbers during phase two do not support the idea that users were motivated to tag videos in VideoTag by a specific interest in the video content, but video content clearly affects tagging behaviour.

8.2.2.4 R4 - Can video tagging games encourage users to enter specific level descriptive tags as well as general level descriptive tags?

There is little research into the quality of the tags that GWAPs produce. This thesis took the view that a good set of tags is one that contains a variety of tags with a wide vocabulary that describes the video content objectively, identifying objects within the video, and subjectively, interpreting the content and offering opinions. Social tags are not useful at describing video content and are generally only meaningful to the tagger or to a select group of users. As summarized above, YouTube tags were found to be mostly social or irrelevant to the video content. Viddler tags were more descriptive of videos than YouTube, but still contained many social and irrelevant tags. The majority of tags entered using VideoTag were descriptive of the video content. Game elements had an effect on tag type which sufficiently changed the types of tag assigned to YouTube videos.

VideoTag users entered tags that identified objects in the video rather than interpreting the content. The same behavior was reported by Gligorov (2012) for users of the Waisida? video tagging game. Tags entered for videos using video tagging games are more likely to be objective at either a basic or specific level rather than perceptual or subjective. In total, 92% of tags entered into VideoTag were descriptive of the video content at either a basic, specific or subjective level, and 88% were objective. The game elements did not encourage many subjective tags, perhaps because tagging at an abstract iconological level is a high cognitive cost activity as suggested by Rafferty and Hilderley (2005). Subjective tags entered were more likely to be expressions of opinion (B3) rather than interpretations of what the content is about (B2). The majority of tags entered into VideoTag were A1 tags that identify objects in the video at a basic level followed by B1b tags that identify people, animals

and inanimate objects at a specific level. The third most entered tag type (B1d) were multiple words or phrases that describe video content using terms that do not appear in existing textual data. There were few subjective tags, but more than social or irrelevant tags. Overall, users tagged videos using specific objective language (50%) more than basic objective language (38%). Although a range of tags is used to describe videos in VideoTag, users use more specific terms. Gligorov *et al.* (2013) states that the combination of basic and specific objective tags is useful for improving textual data for video search. Although users currently use basic terms for video search, if textual data was improved to contain a range of basic and specific terms then more specific searches could be conducted. The fact that users can be encouraged to create specific level tags as well as basic level tags in high quantities, gives scope for further investigations into how user tags can benefit video search.

The large number of A1 and B1b tags highlights an area for improvement in the tag classification model, as also noted by Ransom and Rafferty (2011) and Stuart (2012). The distinction between the two categories is the most difficult to classify and also the most frequently used. The specific level tags are broken down into different categories: places and events (B1a), people, animals and objects (B1b), specific actions (B1c) and multi-words (B1d), whereas basic level descriptions are all categorised as A1. This explains the abundance of A1 tags over any other tag type. A similar breakdown in basic level categories as in specific level categories would allow for a more detailed inspection of how specifically users describe objects in the video. Despite specific level descriptions being split into four categories, the majority of B tags were categorised as B1b. It would be useful to have a further distinction between people, natural objects (e.g., animals, plants, landscape) and inanimate objects (e.g., cars, household objects, buildings). To have more differences coded between A1 and

B1b tag type categories would therefore make the classification more meaningful. Nevertheless, VideoTag was able to encourage users to enter specific level descriptive tags as well as general level descriptive tags.

8.3 Conclusions and Future Work

Overall, this thesis suggests that VideoTag, and therefore video tagging games in general, can help to bridge the gap between automatic methods of indexing and existing text-based search because of the range of descriptive tags produced at both basic and specific levels. VideoTag succeeded in creating a range of descriptive tags, with little evidence of the vocabulary problem and low levels of repetition. Game elements encouraged users to enter tags that described the content but at a more basic level, which could be a result of the one minute time constraint. It would be interesting in future work to remove the time limit from the games to see how the frequencies of the different tag types were affected. Video content was also found to affect the types of tag users enter but there is still a gap in knowledge in whether certain video content can motivate users differently. The phase two experiment suggested that specific interest content could motivate users to enter more tags than could game elements, but it was not clear if it motivated people to use the system. Further investigations into whether specific interest in the video content motivates use are required for more types of content.

This research was limited by low numbers of participants, which is the main limitation. Games With A Purpose are no longer considered novel and this cannot alone motivate use. Moreover, gamification of a video tagging system will probably have limited success unless the non-game system already has a committed user base.

Further research is required into how to motivate users to use video tagging games. Some users who completed the user studies recognised how VideoTag could be useful, but the lack of continued use suggests that it was not useful to them. Low numbers of participants suggests that few users felt the system was useful or enjoyable. Some users felt bored whilst playing so the balance between boredom and anxiety was not designed well enough. There were not enough levels to promote sustained use and users were not adequately supported with new challenges once their skills improved. To attract novice users, goals need to be clearer and learning to play needs to be made part of the fun. New users need to be guided through initial rounds of the games and have tags suggested to them. More playful interactions that improve the juiciness of the interface could be designed to improve appeal. The ability to play on hand held devices would also improve a user's perception of ease of use. However, it is not clear if the game elements are at fault or if the users lacked an interest in the purpose and perceived use of the system. Any improvements to the system and interface design need to be implemented without detracting from the tagging purpose and more consideration needs to be given to how users can benefit from the tags that they enter to encourage them to tag.

8.4 Contribution to Knowledge

The main contributions to knowledge of this thesis are the answers to the four specific research questions, as summarized above. In addition, this research is the first to create video tagging games that do not adopt the two player ESP Game model but use game elements to encourage users to tag videos. It is also the first to apply game theory and motivation theory to GWAP research. The work shows that different elements of gameplay can be designed to encourage users to enter different types of tag. This thesis also offers the first study of how users tag videos that goes

beyond evaluating tag frequency. The research also introduced a custom classification scheme that allowed for the classification of tags entered on YouTube, Viddler and VideoTag. The scheme was proved a reliable method of classification with a Cohen's kappa test for inter-coder reliability giving .831. Finally, the research revealed that users can tag videos using relatively specific language although similar studies have found images to be mainly tagged at a basic level.

9 References

Agresti, A. (2002) Categorical data analysis. John Wiley & Sons.

Al-Khalifa, H. S. and Davis, H. C. (2007) Towards better understanding of folksonomic patterns. Proceedings of the eighteenth conference on Hypertext and hypermedia, pp.163-166. Available from. ISSN 1595938206.

Alexa (2013) The top 500 sites on the web. [online] [Accessed 18 September 2013]. Available at: <<http://www.alexa.com/topsites>>

Ames, M. and Naaman, M. (2007) Why we tag: motivations for annotation in mobile and online media. San Jose, California, USA: ACM, pp. 971-980.

Andrejevic, M. (2009) Exploiting YouTube. in The YouTube Reader. National Library of Sweden., pp.406-423.

Angus, E., Thelwall, M. and Stuart, D. (2008) General patterns of tag usage among university groups in Flickr. Online Information review, 32(1), pp.89-101.

Aparicio, A. F., Vela, F. L. G., Sánchez, J. L. G. and Montes, J. L. I. (2012) Analysis and application of gamification. Proceedings of the 13th International Conference on Interacción Persona-Ordenador, p.17. Available from <<http://dl.acm.org/citation.cfm?id=2379636.2379653>><http://dl.acm.org/ft_gateway.cfm?id=2379653&type=pdf>. ISSN 978-1-4503-1314-8.

Arends, M., Froschauer, J., Goldfarb, D., Merkl, D. and Weingartner, M. (2012) Analysing user motivation in an art folksonomy. Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, p.13. Available from. ISSN 145031242X.

Aurnhammer, M., Hanappe, P. and Steels, L. (2006) Integrating collaborative tagging and emergent semantics for image retrieval. Proc. of the Collaborative Web Tagging Workshop (WWW '06). Available from <<http://www.csl.sony.fr/downloads/papers/2006/aurnhammer-06a.pdf>>.

Bai, L., Lao, S., Zhang, W., Jones, G. F. and Smeaton, A. (2008) Video Semantic Content Analysis Framework Based on Ontology Combined MPEG-7. Adaptive Multimedia Retrieval: Retrieval, User, and Semantics, 4918, pp.237-250.

Bangor, A., Kortum, P. and Miller, J. (2009) Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, 4(3), pp.114-123.

Barendregt, W., Bekker, M. M., Bouwhuis, D. and Baauw, E. (2006) Identifying usability and fun problems in a computer game during first use and after some practice. International Journal of Human-Computer Studies, 64(9), pp.830-846.

Barnum, C. M. and Palmer, L. A. (2010) More than a feeling: understanding the desirability factor in user experience. CHI'10 Extended Abstracts on Human Factors in Computing Systems, pp.4703-4716. Available from. ISSN 1605589306.

Barrington, L., O'malley, D., Turnbull, D. and Lanckriet, G. (2009) User-centered design of a social game to tag music. Paris, France: ACM, pp. 7-10.

Bartle, R. (1996) Hearts, clubs, diamonds, spades: Players who suit MUDs. Journal of MUD research, 1(1), p.19.

Bartle, R. A. (2009) Understanding the Limits of Theory. in Bateman, C. (ed.) Beyond Game Design: Nine Steps to Creating Better Videogames.: Delmar.

Bateman, C., Lowenhaupt, R. and Nacke, L. E. (2011) Player typology in theory and practice. Proceedings of DiGRA. Available from.

Bateman, C. M. and Boon, R. (2005) 21st century game design. Charles River Media Hingham, MA.

Begelman, G., Keller, P. and Smadja, F. (2006) Automated Tag Clustering: Improving search and exploration in the tag space. Collaborative Web Tagging Workshop at WWW2006, pp.15-33. Available from <http://www.pui.ch/phred/automated_tag_clustering/>.

Benedek, J. and Miner, T. (2002) Measuring Desirability: New methods for evaluating desirability in a usability lab setting. Proceedings of Usability Professionals Association, 2003, pp.8-12.

Berendt, B. and Hanser, C. (2007) Tags are not metadata, but "just more content" - to some people. International Conference on Weblogs and Social Media. Available from <<http://www.icwsm.org/papers/paper12.html>>.

Bischoff, K., Firan, C. S., Nejdil, W. and Paiu, R. (2008) Can all tags be used for search? Napa Valley, California, USA: ACM.

Bouca, M. (2012) Mobile communication, gamification and ludification. MindTrek, pp.295-301.

Brooke, J. (1996) SUS-A quick and dirty usability scale. Usability evaluation in industry, 189, p.194.

Brooks, C. and Montanez, N. (2006) Improved annotation of the blogosphere via autotagging and hierarchical clustering. WWW '06: Proceedings of the 15th international conference on World Wide Web, pp.625-632. Available from <<http://portal.acm.org/citation.cfm?id=1135777.1135869>>.

Broxton, T., Interian, Y., Vaver, J. and Wattenhofer, M. (2011) Catching a viral video. *J. Intell. Inf. Syst.*, 40(2), pp.241-259.

Burgess, J. and Green, J. (2009) *YouTube: digital media and society series*. Polity.

Caillois, R. (1961) *Man, play, and games*. University of Illinois Press.

Capocci, A. and Caldarelli, G. (2008) Folksonomies and clustering in the collaborative system CiteULike. *Journal of Physics A: Mathematical and Theoretical*, 41(22).

Capra, R. G., Lee, C. A., Marchionini, G., Russell, T., Shah, C. and Stutzman, F. (2008) Selection and context scoping for digital video collections: an investigation of youtube and blogs. *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pp.211-220.

Carroll, J. M. and Thomas, J. C. (1988) Fun. *ACM SIGCHI Bulletin*, 19(3), pp.21-24.

Carter, M., Gibbs, M. and Harrop, M. (2012) *Metagames, paragames and orthogames: a new vocabulary*. Raleigh, North Carolina: ACM, pp. 11-17.

Cattuto, C., Loreto, V. and Pietronero, L. (2007) Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences (PNAS)*, 104(5), pp.1461-1464.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y. and Moon, S. (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, San Diego, California, USA, pp.1-14. Available from.

Chamberlain, J., Kruschwitz, U. and Poesio, M. (2009) Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. *Proceedings of the 2009 Workshop on The People's Web Meets NLP:*

Collaboratively Constructed Semantic Resources, pp.57-62. Available from. ISSN 1932432558.

Chamberlain, J., Kruschwitz, U. and Poesio, M. (2012) Motivations for participation in socially networked collective intelligence systems. arXiv preprint arXiv:1204.4071.

Chen, J. (2007) Flow in games (and everything else). *Communications of the ACM*, 50(4), pp.31-34.

Chen, M., Kolko, B. E., Cuddihy, E. and Medina, E. (2011) Modeling but NOT measuring engagement in computer games. *Proceedings of the 7th international conference on Games + Learning + Society Conference*, pp.55-63. Available from <<http://dl.acm.org/citation.cfm?id=2206376.2206383>><http://dl.acm.org/ft_gateway.cfm?id=2206383&type=pdf>.

Cheng, X., Dale, C. and Liu, J. (2008) Statistics and Social Network of YouTube Videos. *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pp.229-238. Available from. ISSN 1548-615X.

Cherry, M. A. (2012) The Gamification of Work. *HOFSTRA LAW REVIEW*, 40, p.851.

Chi, E. and Mytkowicz, T. (2007) Understanding Navigability of Social Tagging Systems. *Computer/Human Interaction*. Available from <<http://www.viktoria.se/altchi/index.php?action=showsubmission&id=39>>.

Chi, E. H. and Mytkowicz, T. (2008) Understanding the efficiency of social tagging systems using information theory. *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pp.81-88. Available from. ISSN 1595939857.

Chiang, Y.-T., Cheng, C.-Y. and Lin, S. S. J. (2008) The effects of digital games on undergraduate players' flow experiences and affect. *Digital Games and Intelligent*

Toys Based Education, 2008 Second IEEE International Conference on, pp.157-159.
Available from.

Chorney, A. I. (2013) Taking the game out of gamification. *Dalhousie Journal of Interdisciplinary Management*, 8(1).

Cohn, J. P. (2008) Citizen science: Can volunteers do real research? *BioScience*, 58(3), pp.192-197.

Conduit, N. and Rafferty, P. (2007) Constructing an image indexing template for The Children's Society: Users' queries and archivists' practice. *Journal of Documentation*, 63(6), pp.898-919.

Cortina, J. M. (1993) What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, 78(1), p.98.

Cowley, B., Charles, D., Black, M. and Hickey, R. (2008) Toward an understanding of flow in video games. *Computers in Entertainment (CIE)*, 6(2).

Croft, W. and Cruse, D. A. (2004) *Cognitive linguistics*. Cambridge University Press.

Csikszentmihalyi, M. (1975) Play and intrinsic rewards. *Journal of humanistic psychology*.

Csikszentmihalyi, M. (1997) *Finding flow: The psychology of engagement with everyday life*. Basic Books.

Csikszentmihalyi, M. (2000) *Beyond boredom and anxiety*. Jossey-Bass.

Csikszentmihalyi, M. and Lefevre, J. (1989) Optimal experience in work and leisure. *Journal of personality and social psychology*, 56(5).

Cunningham, S. J. and Nichols, D. M. (2008) How people find videos. Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, Pittsburgh PA, PA, USA, pp.201-210. Available from <<http://doi.acm.org/10.1145/1378889.1378924>>. ISSN 978-1-59593-998-2.

De Rooij, O., Snoek, C. G. M. and Worring, M. (2008) Balancing thread based navigation for targeted video search. Proceedings of the 2008 international conference on Content-based image and video retrieval, Niagara Falls, Canada, pp.485-494. Available from <<http://doi.acm.org/10.1145/1386352.1386414>>. ISSN 978-1-60558-070-8.

Desurvire, H., Caplan, M. and Toth, J. A. (2004) Using heuristics to evaluate the playability of games. Vienna, Austria: ACM, pp. 1509-1512.

Deterding, S., Dixon, D., Khaled, R. and Nacke, L. (2011a) From game design elements to gamefulness: defining gamification. Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp.9-15. Available from <<http://dl.acm.org/citation.cfm?id=2181037.2181040>><http://dl.acm.org/ft_gateway.cfm?id=2181040&type=pdf>. ISSN 978-1-4503-0816-8.

Deterding, S., Khaled, R., Nacke, L. E. and Dixon, D. (2011b) Gamification: Toward a definition. CHI 2011 Gamification Workshop Proceedings.

Ding, Y., Jacob, E. K., Zhang, Z., Foo, S., Yan, E., George, N. L. and Guo, L. (2009) Perspectives on social tagging. Journal of the American Society for Information Science and Technology, 60(12), pp.2388-2401.

Dixon, D. (2011) Player types and gamification. Proceedings of the CHI 2011 Workshop on Gamification.

- Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P. and Tomkins, A. (2007) Visualizing tags over time. *ACM Transactions on the Web (TWEB)*, 1(2).
- Dunn, S. and Hedges, M. (2012) *Crowd-Sourcing Scoping Study-Engaging the Crowd with Humanities Research*. online repository: Centre for e-Research, King's College London.
- Enser, P. (2008) The evolution of visual information retrieval. *Journal of Information Science*.
- Enser, P. G., Sandom, C. J., Hare, J. S. and Lewis, P. H. (2007) Facing the reality of semantic image retrieval. *Journal of Documentation*, 63(4), pp.465-481.
- Experian Online Video: Bringing Social Media to Life. (2011) Experian Hitwise. [online] Available at <<http://www.hitwise.com/uk/registration-pages/online-video-bringing-social-media-to-life/>>
- Febretti, A. and Garzotto, F. (2009) Usability, playability, and long-term engagement in computer games. Boston, MA, USA: ACM, pp. 4063-4068.
- Federoff, M. A. (2002) Heuristics and usability guidelines for the creation and evaluation of fun in video games. Master of Science. Theses, Indiana University, Bloomington.
- Field, A. (2013) *Discovering statistics using IBM SPSS statistics*. Sage.
- Figueiredo, F., Benevenuto, F. I. and Almeida, J. (2011) The Tube Over Time: Characterizing Popularity Growth of YouTube Videos. *Proceedings of the 4th ACM International Conference of Web Search and Data Mining (WSDM'11)*.
- Fine, G. A. (1987) Meaningful play, playful meaning. *Human Kinetics*.

Fleiss, J. L., Levin, B. and Paik, M. C. (2013) Statistical methods for rates and proportions. John Wiley & Sons.

Fortugno, N. (2008) The strange case of casual games. *Game Usability: Advancing the Player Experience* (K. Isbister and N. Schaffer, Eds.). Burlington: Elsevier, pp.143-159.

Freiburg, B., Kamps, J. and Snoek, C. G. (2011) Crowdsourcing visual detectors for video search. *Proceedings of the 19th ACM international conference on Multimedia*, pp.913-916. Available from. ISSN 1450306160.

Fullerton, T. (2008) *Game design workshop: a playcentric approach to creating innovative games*. Taylor & Francis US.

Furnas, G., Fake, C., Von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C. and Naaman, M. (2006) Why do tagging systems work? *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pp.36-39. Available from <<http://portal.acm.org/citation.cfm?id=1125451.1125462>>.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987) The vocabulary problem in human-system communication. *Commun. ACM*, 30(11), pp.964-971.

Geisler, G. and Burns, S. (2007) *Tagging video: conventions and strategies of the YouTube community*. Vancouver, BC, Canada: ACM, pp. 480-480.

Gligorov, R. (2012) User-generated metadata in audio-visual collections. *Proceedings of the 21st international conference companion on World Wide Web*, pp.139-144. Available from <http://dl.acm.org/ft_gateway.cfm?id=2187998&type=pdf>.

Gligorov, R., Baltussen, L. B., Van Ossenbruggen, J., Aroyo, L., Brinkerink, M., Oomen, J. and Van Ees, A. (2010) Towards Integration of End-User Tags with Professional Annotations. Proceedings of the WebSci10: Extending the Frontiers of Society On-Line. Available from <<http://journal.webscience.org/363/>>.

Gligorov, R., Hildebrand, M., Van Ossenbruggen, J., Aroyo, L. and Schreiber, G. (2013) An evaluation of labelling-game data for video retrieval. Proceedings of the 35th European conference on Advances in Information Retrieval, pp.50-61. Available from <<http://dl.acm.org/citation.cfm?id=2458181.2458188>>. ISSN 978-3-642-36972-8.

Gligorov, R., Hildebrand, M., Van Ossenbruggen, J., Schreiber, G. and Aroyo, L. (2011) On the role of user-generated metadata in audio visual collections. Banff, Alberta, Canada: ACM, pp. 145-152.

Goh, D. H.-L., Ang, R. P., Chua, A. Y. K. and Lee, C. S. (2010) Evaluating game genres for tagging images. Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp.659-662. Available from <<http://dl.acm.org/citation.cfm?id=1868914.1868998>><http://dl.acm.org/ft_gateway.cfm?id=1868998&type=pdf>. ISSN 978-1-60558-934-3.

Goh, D. H. L., Ang, R. P., Lee, C. S. and Chua, A. Y. K. (2011) Fight or unite: Investigating game genres for image tagging. Journal of the American Society for Information Science and Technology, 62(7), pp.1311-1324.

Golder, S. A. and Huberman, B. A. (2005) The structure of collaborative tagging system. Information Dynamics Lab: HP Labs, Palo Alto, USA.

Golder, S. and Huberman, B. (2006) Usage patterns of collaborative tagging systems. J. Inf. Sci., 32(2), pp.198-208.

Gomes, J. M., Chambel, T. and Langlois, T. (2013a) Engaging Users in Audio Labelling as a Movie Browsing Game with a Purpose. in *Advances in Computer Entertainment*. Springer, pp.296-307.

Gomes, J. M. A., Chambel, T. and Langlois, T. (2013b) SoundsLike: movies soundtrack browsing and labeling based on relevance feedback and gamification. Como, Italy: ACM, pp. 59-62.

Goodrum, A. (2003) If it sounds as good as it looks: Lessons learned from video retrieval evaluation. *The MIR/MDL Evaluation Project White Paper Collection*, 3, pp.97-102.

Graham, R. and Caverlee, J. (2008) Exploring Feedback Models in Interactive Tagging. 1 *IEEE Computer Society*, pp. 141-147.

Greenaway, S. (2007) VideoTag: A game to encourage tagging of videos to improve accessibility and search. *MSc Computer Science (Internet Technology)*. Theses, University of Wolverhampton.

Groh, F. (2012) Gamification: State of the art definition and utilization. *Institute of Media Informatics Ulm University*, pp.39-47.

Guy, M. and Tonkin, E. (2006) Tidying up tags. *D-lib Magazine*, 12(1), pp.1082-9873.

Hare, J. S., Lewis, P. H., Enser, P. G. and Sandom, C. J. (2006) Mind the gap: another look at the problem of the semantic gap in image retrieval. *Electronic Imaging 2006*, pp. 309-312.

Halpin, H., Robu, V. and Shepherd, H. (2007) The complex dynamics of collaborative tagging. Banff, Alberta, Canada: ACM, pp. 211-220.

Halvey, M. and Jose, J. (2012) Bridging the gap between expert and novice users for video search. *International Journal of Multimedia Information Retrieval*, 1(1), pp.17-29.

Halvey, M. J. and Keane, M. T. (2007) Analysis of online video search and sharing. *Proceedings of the eighteenth conference on Hypertext and hypermedia*, Manchester, UK, pp.217-226.

Hamari, J. (2013) Transforming Homo Economicus into Homo Ludens: a field experiment on gamification in a utilitarian peer-to-peer trading service. *Electronic Commerce Research and Applications*.

Hamari, J. and Koivisto, J. (2013) Social motivations to use gamification: An empirical study of gamifying exercise. *proceedings of the 21 st European conference in information systems*. Utrecht, Netherlands.

Hamari, J., Koivisto, J. and Sarsa, H. (2014) Does Gamification Work? — A Literature Review of Empirical Studies on Gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences*. HICSS.

Hildebrand, M., Brinkerink, M., Gligorov, R., Steenbergen, M. V., Huijkman, J. and Oomen, J. (2013) Waisda?: video labeling game. *Barcelona, Spain: ACM*, pp. 823-826.

Hildebrand, M. and Ossenbruggen, J. V. (2012) Linking user generated video annotations to the web of data. *Klagenfurt, Austria: Springer-Verlag*, pp. 693-704.

Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J. Y.-J. and Chen, K.-T. (2009) KissKissBan: a competitive human computation game for image annotation. *Paris, France: ACM*, pp. 11-14.

Hollink, L., Schreiber, A. T., Wielinga, B. J. and Worrying, M. (2004) Classification of user image descriptions. *International Journal of Human-Computer Studies*, 61(5), pp.601-626.

Hsu, C.-L. and Lu, H.-P. (2004) Why do people play on-line games? an extended TAM with social influences and flow experience. *Inf. Manage.*, 41(7), pp.853-868.

Hsu, Y. P., Hsu, K. C. and Ko, C. M. (2007) Casual games provide stress relief for female gamer in Taiwan. *Nr Reading: Academic Conferences Ltd.*

Huang, J., Thornton, K. M. and Efthimiadis, E. N. (2010) Conversational tagging in twitter. *Toronto, Ontario, Canada: ACM*, pp. 173-178.

Huizinga, J. (1949) *Homo ludens*. Taylor & Francis.

Hunicke, R., Leblanc, M. and Zubek, R. (2004) *MDA: A Formal Approach to Game Design and Game Research*. AAAI Press.

Huotari, K. and Hamari, J. (2012) Defining gamification: a service marketing perspective. *Proceeding of the 16th International Academic MindTrek Conference*, pp.17-22. Available from <http://dl.acm.org/citation.cfm?id=2393132.2393137> <http://dl.acm.org/ft_gateway.cfm?id=2393137&type=pdf>. ISSN 978-1-4503-1637-8.

Iacovides, I., Jennett, C., Cornish-Trestrail, C. and Cox, A. L. (2013) Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp.1101-1106. Available from. ISSN 1450319521.

Isbister, K. (2010) Enabling social play: A framework for design and evaluation. in *Evaluating User Experience in Games*. Springer, pp.11-22.

- Iso, I. O. F. S. (1998) ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability.
- Jaimes, A. and Chang, S.-F. (2000) A conceptual framework for indexing visual information at multiple levels. *IS&T/SPIE Internet Imaging*, 3964, pp.2-15.
- Jaimes, A., Tseng, B. and Smith, J. (2003) Modal Keywords, Ontologies, and Reasoning for Video Understanding. *Image and Video Retrieval*, 2728, pp.248-259.
- Jain, S. and Parkes, D. C. (2009) The role of game theory in human computation systems. Paris, France: ACM, pp. 58-61.
- Jegers, K. (2007) Pervasive game flow: understanding player enjoyment in pervasive gaming. *Computers in Entertainment (CIE)*, 5(1), p.9.
- Johnson, F. (2012) Using semantic differentials for an evaluative view of the search engine as an interactive system. Nijmegen.
- Juul, J. (2009) *A Casual Revolution*. The MIT Press.
- Jørgensen, C., Jaimes, A., Benitez, A. B. and Chang, S. F. (2001) A conceptual framework and empirical research for classifying visual descriptors. *Journal of the American Society for Information Science and Technology*, 52(11), pp.938-947.
- Kallio, K. P., Mäyrä, F. and Kaipainen, K. (2011) At least nine ways to play: approaching gamer mentalities. *Games and Culture*, 6(4), pp.327-353.
- Kern, R., Granitzer, M. and Pammer, V. (2008) Extending folksonomies for image tagging. *Image Analysis for Multimedia Interactive Services*, 2008. WIAMIS'08. Ninth International Workshop on, pp.126-129. Available from. ISSN 0769533442.

- Kim, H. H. (2011) Toward video semantic search based on a structured folksonomy. *Journal of the American Society for Information Science and Technology*, 62(3), pp.478-492.
- Kim, J. (2012) The institutionalization of YouTube: From user-generated content to professionally generated content. *Media, Culture & Society*, 34(1), pp.53-67.
- Kim, Y. E., Schmidt, E. M. and Emelle, L. (2008) MoodSwings: A Collaborative Game for Music Mood Label Collection. *ISMIR*, 8, pp.231-236.
- Kipp, M. E. and Campbell, D. G. (2006) Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology*, 43(1), pp.1-18.
- Kipp, M. E. I. (2006) Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science*, 30(3).
- Kipp, M. E. I. (2007a) Tagging Practices on Research Oriented Social Bookmarking Sites. 35th conference of the Canadian Association for Information Science.
- Kipp, M. E. I. (2007b) @toread and Cool: Tagging for Time, Task and Emotion. 8th Information Architecture Summit. Available from <<http://dlist.sir.arizona.edu/1947/>>.
- Kofler, C., Yang, L., Larson, M., Mei, T., Hanjalic, A. and Li, S. (2012) When video search goes wrong: predicting query failure using search engine logs and visual search results. Nara, Japan: ACM, pp. 319-328.
- Kowal, J. and Fortier, M. S. (1999) Motivational Determinants of Flow: Contributions From Self-Determination Theory. *The Journal of Social Psychology*, 139(3), pp.355-368.

Kuittinen, J., Kultima, A., Niemelä, J., and Paavilainen, J. (2007) Casual games discussion. Toronto, Canada: ACM, pp. 105-112.

Körner, C., Benz, D., Hotho, A., Strohmaier, M. and Stumme, G. (2010) Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. Proceedings of the 19th international conference on World wide web, pp.521-530. Available from. ISSN 1605587990.

Laitinen, S. (2006) Do usability expert evaluation and test provide novel and useful data for game development. Journal of usability studies, 2(1), pp.64-75.

Laitinen, S. (2008) Usability and playability expert evaluation. Game Usability: Advice From the Experts for Advancing the Player Experience. Morgan Kaufmann. Burlington, MA, USA, pp.91-112.

Lambiotte, R. and Ausloos, M. (2006) Collaborative tagging as a tripartite network. in Computational Science–ICCS 2006. Springer, pp.1114-1117.

Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. biometrics, 33(1), pp.159-174.

Law, E. and Von Ahn, L. (2009) Input-agreement: A New Mechanism for Data Collection using Human Computation Games. CHI, pp.1197-1206.

Lazzaro, N. (2008) The four fun keys. Game Usability: Advancing the Player Experience (K. Isbister and N. Schaffer, Eds.). Burlington: Elsevier, pp.315-344.

Lee, J. and Hwang, S.-W. (2008) Ranking with tagging as quality indicators. Proceedings of the 2008 ACM symposium on Applied computing, pp.2432-2436. Available from. ISSN 1595937536.

Leyssen, M. H. R., Traub, M. C., Van Ossenbruggen, J. R. and Hardman, L. (2012) Is It A Bird Or Is It A Crow? The Influence Of Presented Tags On Image Tagging By Non-Expert Users. CWI, CWI Tech.Report INS-1202.

Li, H., Liu, J., Xu, K. and Wen, S. (2012) Understanding video propagation in online social networks. Coimbra, Portugal: IEEE Press, pp. 1-9.

Lin, C.-W., Chen, K.-T., Chen, L.-J., King, I. and Lee, J. H.-M. (2008) An analytical approach to optimizing the utility of ESP games. Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, 1, pp.184-187. Available from. ISSN 0769534961.

Lin, X., Beaudoin, J. E., Bui, Y. and Desai, K. (2006) Exploring characteristics of social classification. Advances in Classification Research Online, 17(1), pp.1-19.

Lin, Y.-L. and Aroyo, L. (2012) Interactive curating of user tags for audiovisual archives. Proceedings of the International Working Conference on Advanced Visual Interfaces, pp.685-688. Available from <<http://dl.acm.org/citation.cfm?id=2254556.2254685>><http://dl.acm.org/ft_gateway.cfm?id=2254685&type=pdf>. ISSN 978-1-4503-1287-5.

Liu, Y., Alexandrova, T. and Nakajima, T. (2011) Gamifying intelligent environments. Proceedings of the 2011 international ACM workshop on Ubiquitous meta user interfaces, pp.7-12. Available from. ISSN 1450309933.

Loui, A., Luo, J., Chang, S.-F., Ellis, D., Jiang, W., Kennedy, L., Lee, K. and Yanagawa, A. (2007) Kodak's consumer video benchmark data set: concept definition and annotation. Proceedings of the international workshop on Workshop on multimedia information retrieval, Augsburg, Bavaria, Germany, pp.245-254. Available from <<http://doi.acm.org/10.1145/1290082.1290117>>. ISSN 978-1-59593-778-0.

- Ma, H., Chandrasekar, R., Quirk, C. and Gupta, A. (2009) Improving search engines using human computation games. Proceeding of the 18th ACM conference on Information and knowledge management, Hong Kong, China, pp.275-284.
- Macarthur, A. (2014) Where Hashtags Really Came From and What They Do. [Accessed 22 March 2014]. Available at: <<http://twitter.about.com/od/Twitter-Hashtags/a/The-History-Of-Hashtags.htm>>.
- Macgregor, G. and Mcculloch, E. (2006) Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5), pp.291-300.
- Malone, T. W. (1982) Heuristics for designing enjoyable user interfaces: Lessons from computer games. Proceedings of the 1982 conference on Human factors in computing systems, pp.63-68. Available from.
- Mandel, M. I. and Ellis, D. P. (2008) A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2), pp.151-165.
- Mannell, R. C., Zuzanek, J. and Larson, R. (1988) Leisure states and "flow" experiences: Testing perceived freedom and intrinsic motivation hypotheses. *Journal of Leisure Research*.
- Marchionini, G., Shah, C., Lee, C. A. and Capra, R. (2009) Query parameters for harvesting digital video and associated contextual information. Austin, TX, USA: ACM, pp. 77-86.
- Marlow, C., Naaman, M., Boyd, D. and Davis, M. (2006) HT06, tagging paper, taxonomy, Flickr, academic article, to read. *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pp.31-40. Available from <<http://portal.acm.org/citation.cfm?id=1149949>>.

Mason, W. and Watts, D. J. (2010) Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2), pp.100-108.

Massimini, F., Csikszentmihalyi, M. and Fave, A. D. (1988) Flow and biocultural evolution.

Mathes, A. (2004) *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*.

Mcdonald, S. and Tait, J. (2003) Search strategies in content-based image retrieval. Toronto, Canada: ACM, pp. 80-87.

Mcgonigal, J. (2011) *Reality is broken: Why games make us better and how they can change the world*. Penguin. com.

Mekler, E. D., Brühlmann, F., Opwis, K. and Tuch, A. N. (2013) Disassembling gamification: the effects of points and meaning on user motivation and performance. *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp.1137-1142. Available from. ISSN 1450319521.

Melenhorst, M., Grootveld, M., Setten, M. V. and Veenstra, M. (2008) Tag-based information retrieval of video content. Silicon Valley, California, USA: ACM, pp. 31-40.

Melenhorst, M. and Van Velsen, L. (2010) TEMPTING TO TAG: AN EXPERIMENTAL COMPARISON OF FOUR TAGGING INPUT MECHANISMS. *Human Technology*, 6(2).

Merholz, P. (2005) peterme.com: Clay Shirky's Viewpoints are Overrated.

Min, X., Maddage, N. C., Changsheng, X., Kankanhalli, M. and Qi, T. (2003) Creating audio keywords for event detection in soccer video. *Multimedia and Expo, 2003*.

ICME '03. Proceedings. 2003 International Conference on, 2, pp.II-281-284 vol.282. Available from.

Morsillo, N., Mann, G. and Pal, C. (2010) YouTube Scale, Large Vocabulary Video Annotation. Video Search and Mining, 287, pp.357-386.

Müller, A., Lux, M. and Böszörmenyi, L. (2012) The video summary GWAP: summarization of videos based on a social game. Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, p.15. Available from <http://dl.acm.org/citation.cfm?id=2362456.2362476> <http://dl.acm.org/ft_gateway.cfm?id=2362476&type=pdf>. ISSN 978-1-4503-1242-4.

Neomobile (2013) Instagram Video Vs. Vine - Neomobile Blog. 2013. {blog entry}. [Accessed 12 November 2013]

Newman, M. E. J. (2005) Power laws, Pareto distributions and Zipf's law. Contemporary physics, 46(5), pp.323-351.

Nicholson, S. (2012) A user-centered theoretical framework for meaningful gamification. Proceedings GLS, 8.

Nielsen, J. (1994) Usability engineering. Elsevier.

Nielsen, J. and Molich, R. (1990) Heuristic evaluation of user interfaces. Proceedings of the SIGCHI conference on Human factors in computing systems, pp.249-256. Available from. ISSN 0201509326.

Nunnally, J. (1978) C.(1978). Psychometric theory. New York: McGraw-Hill.

O'brien, H. L. and Toms, E. G. (2010) Is there a universal instrument for measuring interactive information retrieval?: the case of the user engagement scale. New Brunswick, New Jersey, USA: ACM, pp. 335-340.

Olson, C. L. (1976) On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4), p.579.

Oomen, J., Belice Baltussen, L., Limonard, S., Van Ees, A., Brinkerink, M., Aroyo, L., Vervaart, J., Asaf, K. and Gligorov, R. (2010) Emerging practices in the cultural heritage domain-social tagging of audiovisual heritage. Web Science Trust.

Orehovacki, T. (2010) Proposal for a set of quality attributes relevant for Web 2.0 application success. *Information Technology Interfaces (ITI)*, 2010 32nd International Conference on, pp.319-326. Available from. ISSN 1424457327.

Over, P. (2014) TRECVID 2013 Guidelines. [online]. [Accessed 3 November 2013]. Available at: <<http://www-nlpir.nist.gov/projects/tv2013/>>

Panofsky, E. (1970) *Meaning in the visual arts*. Penguin books Harmondsworth.

Paolillo, J. C. and Penumarthy, S. (2007) The Social Structure of Tagging Internet Video on del.icio.us. *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pp.85-85. Available from <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4076541>.

Passant, A. (2007) Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs: Theoretical background and corporate use-case. *International Conference on Weblogs and Social Media*. Available from <<http://www.icwsm.org/papers/paper15.html>>.

Pillai, K. (1998) Pillai's Trace. *Encyclopedia of statistical sciences*.

Pinto, J. P. and Viana, P. (2013) TAG4VD: a game for collaborative video annotation. Barcelona, Spain: ACM, pp. 25-28.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L. and Ducceschi, L. (2013) Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1), pp.1-44.

Rafelsberger, W. and Scharl, A. (2009) Games with a purpose for social networking platforms. *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pp.193-198.

Rafferty, P. (2011) Informative Tagging of Images: The Importance of Modality in Interpretation. *Knowledge organization*, 38(4).

Rafferty, P. and Hilderley, R. (2005) Indexing multimedia and creative works: the problems of meaning and interpretation. Ashgate Publishing, Ltd.

Rafferty, P. and Hilderley, R. (2007) Flickr and Democratic Indexing: dialogic approaches to indexing. *Aslib Proceedings*, 59(4/5), pp.397 - 410.

Rainie, L. (2014) Three Technology Revolutions. Available From <www.pewinternet.org>.

Ransom, N. and Rafferty, P. (2011) Facets of user-assigned tags and their effectiveness in image retrieval. *Journal of Documentation*, 67(6), pp.1038-1066.

Reimer, C. (2011) Play to order: what Huizinga has to say about gamification. *Proceedings of the 7th international conference on Games + Learning + Society Conference*, pp.272-274. Available from <<http://dl.acm.org/citation.cfm?id=2206376.2206414>><http://dl.acm.org/ft_gateway.cfm?id=2206414&type=pdf>.

- Robertson, S., Vojnovic, M. and Weber, I. (2009) Rethinking the ESP game. Boston, MA, USA: ACM, pp. 3937-3942.
- Rorissa, A. (2008) User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. *Information Processing & Management*, 44(5), pp.1741-1753.
- Rosch, E. (1975) Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), p.192.
- Rotman, D. and Preece, J. (2010) The 'WeTube' in YouTube – creating an online community through video sharing. *Int.J.Web Based Communities*, 6(3), pp.317-333.
- Ryan, R. M. and Deci, E. L. (2000) Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 25(1), pp.54-67.
- Salen, K. and Zimmerman, E. (2004) *Rules of play: game design fundamentals*. 2004. Massachusetts Institute of Technology.
- Sauro, J. (2011) Measuring usability with the system usability scale (SUS). 2012-06-25[2011-05-08].[Http://www.measuringusability, com/sus.php](http://www.measuringusability.com/sus.php).
- Sauro, J. and Lewis, J. R. (2011) When designing usability questionnaires, does it hurt to be positive? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.2215-2224. Available from. ISSN 1450302289.
- Schaffer, N. (2008) Heuristic evaluation of games. *Game Usability*.
- Schmitz, P. (2006) Inducing ontology from flickr tags. *Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland.

Sen, S., Harper, F. M., Lapitz, A. and Riedl, J. (2007) The quest for quality tags. Sanibel Island, Florida, USA: ACM.

Sen, S., Lam, S., Rashid, A., Cosley, D., Frankowski, D., Osterhouse, J., Harper, M. and Riedl, J. (2006) tagging, communities, vocabulary, evolution. CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, pp.181-190. Available from <<http://portal.acm.org/citation.cfm?id=1180875.1180904>>.

Shamma, D. A., Shaw, R., Shafton, P. L. and Liu, Y. (2007) Watch what I watch: using community activity to understand content. Augsburg, Bavaria, Germany: ACM, pp. 275-284.

Shatford, S. (1986) Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly*, 6(3), pp.39-62.

Sheng, V. S., Provost, F. and Ipeirotis, P. G. (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.614-622. Available from. ISSN 1605581933.

Shih-Fu, C., Wei-Ying, M. and Smeulders, A. (2007) Recent Advances and Challenges of Semantic Image/Video Search. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 4, pp.-1205-1208. Available from. ISSN 1520-6149.

Shin, D.-H. and Shin, Y.-J. (2011) Why do people play social network games? *Computers in Human Behavior*, 27(2), pp.852-861.

Shirky, C. (2005) *Ontology is Overrated -- Categories, Links, and Tags*.

Shneiderman, B. (1987) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Publ. Co., Reading, MA.

Shneiderman, B. (2004) Designing for fun: how can we design user interfaces to be more fun? *interactions*, 11(5), pp.48-50.

Silva, P. A. and Dix, A. (2007) *Usability: not as we know it!* University of Lancaster, United Kingdom: British Computer Society, pp. 103-106.

Sinha, R. (2005) A cognitive analysis of tagging.

Siorpaes, K. and Hepp, M. (2008) *OntoGame: Weaving the Semantic Web by Online Games*. in Bechhofer, S., Hauswirth, M., Hoffmann, J. and Koubarakis, M. (eds.) *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, pp.751-766.

Smeaton, A. F., Over, P. and Kraaij, W. (2009) High-Level Feature Detection from Video in TRECVID: a 5-Year Retrospective of Achievements. in Ajay, D. (ed.) *Multimedia Content Analysis, Theory and Applications*. Berlin: Springer Verlag, pp.151-174.

Sood, S., Owsley, S., Hammond, K. and Birnbaum, L. (2007) *TagAssist: Automatic Tag Suggestion for Blog Posts*. International Conference on Weblogs and Social Media. Available from <<http://www.icwsm.org/papers/paper10.html>>.

Strohmaier, M., Körner, C. and Kern, R. (2012) Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17, pp.1-11.

Stuart, E. (2012) *Motivations to Upload and Tag Images vs. Tagging Practice: An Investigation of the Web 2.0 Site Flickr*. PhD. Theses, University of Wolverhampton.

Sutton-Smith, B. (1997) *The ambiguity of play*. Harvard University Press.

- Sweetser, P. and Wyeth, P. (2005) GameFlow: a model for evaluating player enjoyment in games. *Comput. Entertain.*, 3(3), pp.3-3.
- Tao, D., Adsul, P., Wray, R., Jupka, K., Semar, C. and Goggins, K. (2012) Search strategy effectiveness and relevance of YouTube videos. *Proceedings of the American Society for Information Science and Technology*, 49(1), pp.1-4.
- Thaler, S., Siorpaes, K., Simperl, E. and Hofer, C. (2011) A survey on games for knowledge acquisition. *Rapport technique, STI*.
- Tjondronegoro, D. and Spink, A. (2008) Web search engine multimedia functionality. *Inf.Process.Manage.*, 44(1), pp.340-357.
- Tjondronegoro, D., Spink, A. and Jansen, B. J. (2009) A study and comparison of multimedia Web searching: 1997–2006. *Journal of the American Society for Information Science and Technology*, 60(9), pp.1756-1768.
- Trant, J. (2006) Social Classification and Folksonomy in Art Museums: early data from the steve.museum tagger prototype. ASIST SIG-CR workshop on Social Classification, November 4, 2006. Available from <<http://www.archimuse.com/papers/asist-CR-steve-0611.pdf>>.
- Trant, J. (2009) Tagging, Folksonomy and Art Museums: Results of steve.museum's research.
- Trant, J. and Wyman, B. (2006) Investigating social tagging and folksonomy in art museums with steve.museum. In proceedings of: WWW'06 Collaborative Web Tagging Workshop.
- Tullis, T. S. and Stetson, J. N. (2004) A comparison of questionnaires for assessing website usability. Usability Professional Association Conference.

- Turnbull, D., Liu, R., Barrington, L. and Lanckriet, G. R. (2007) A Game-Based Approach for Collecting Semantic Annotations of Music. *ISMIR*, 7, pp.535-538.
- Turner, P. (2010) The anatomy of engagement. In proceedings of the 28th Annual European Conference on Cognitive Ergonomics, pp.59-66.
- Tuunanen, J. and Hamari, J. (2012) Meta-synthesis of player typologies. In proceedings of Nordic Digra 2012 Conference: Games in Culture and Society, Tampere, Finland.
- Ulges, A., Borth, D. and Koch, M. (2013) Content analysis meets viewers: linking concept detection with demographics on YouTube. *International Journal of Multimedia Information Retrieval*, 2(2), pp.145-157.
- Ulges, A., Schulze, C., Keysers, D. and Breuel, T. (2008a) A System that Learns to Tag Videos by Watching Youtube. *Int. Conf. on Vision Systems (ICVS)*, pp.415-424.
- Ulges, A., Schulze, C., Keysers, D. and Breuel, T. (2008b) Identifying relevant frames in weakly labeled videos for training concept detectors. Niagara Falls, Canada: ACM, pp. 9-16.
- Vallerand, R. J. (1997) Toward A Hierarchical Model of Intrinsic and Extrinsic Motivation. in Mark, P. Z. (ed.) *Advances in Experimental Social Psychology*. Academic Press, pp.271-360.
- Vallet, D., Hopfgartner, F., Halvey, M. and Jose, J. (2008) Community based feedback techniques to improve video search. *Signal, Image and Video Processing*, 2(4), pp.289-306.
- Van Velsen, L. and Melenhorst, M. (2009) Incorporating user motivations to design for video tagging. *Interacting with Computers*, 21(3), p.221-232.

- Vander Wal, T. (2007) Folksonomy coinage and definition.
- Velsen, L. V. and Melenhorst, M. (2009) Incorporating user motivations to design for video tagging. *Interact. Comput.*, 21(3), pp.221-232.
- Venhuizen, N. J., Basile, V., Evang, K. and Bos, J. (2013) Gamification for word sense labeling. *Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pp.397-403.
- Voiskounsky, A. E., Mitina, O. V. and Avetisova, A. A. (2004) Playing Online Games: Flow Experience. *PsychNology journal*, 2(3), pp.259-281.
- Von Ahn, L. (2005) Human Computation. PhD. Theses, Carnegie Mellon.
- Von Ahn, L. (2006) Games with a Purpose. *Computer*, 39(6), pp.92-94.
- Von Ahn, L. and Dabbish, L. (2004) Labeling images with a computer game. *CHI '04: Proceedings of the 2004 conference on Human factors in computing systems*, pp.319-326. Available from <<http://portal.acm.org/citation.cfm?id=985692.985733>>.
- Von Ahn, L. and Dabbish, L. (2008) Designing games with a purpose. *Commun. ACM*, 51(8), pp.58-67.
- Von Ahn, L., Liu, R. and Blum, M. (2006) Peekaboom: a game for locating objects in images. *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp.55-64. Available from. ISSN 1595933727.
- Voss, J. (2007) Tagging, folksonomy & co-renaissance of manual indexing?
- Wang, M., Ni, B., Hua, X.-S. and Chua, T.-S. (2012) Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput.Surv.*, 44(4), pp.25:21-25:24.

Wikipedia (2014) 1% rule (Internet culture) [online]. [Accessed 18 April 2014]. Available at: <http://en.wikipedia.org/wiki/1%25_rule_%28Internet_culture%29>

Williams, D., Kelly, G. and Anderson, L. (2004) MSN 9: new user-centered desirability methods produce compelling visual design. CHI'04 Extended Abstracts on Human Factors in Computing Systems, pp.959-974. Available from. ISSN 1581137036.

Wolfson, S. and Case, G. (2000) The effects of sound and colour on responses to a computer game. *Interacting with computers*, 13(2), pp.183-192.

Wu, H., Zubair, M. and Maly, K. (2006) Harvesting social knowledge from folksonomies. *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pp.111-114. Available from <<http://portal.acm.org/citation.cfm?id=1149962>>.

Yee, N. (2006) Motivations for play in online games. *CyberPsychology & Behavior*, 9(6), pp.772-775.

Youtube (2013) Statistics. [Accessed 12 September 2013]. Available at: <<http://www.youtube.com/yt/press/statistics.html>>

Yu, C. H. (2001) An introduction to computing and interpreting Cronbach coefficient alpha in SAS. *Proceedings of 26th SAS User Group International Conference*, pp.22-25.

Appendix A

Research in progress paper published and presented at ISSI 2009.

Tagging YouTube - A Classification of Tagging Practice on YouTube

Stacey Greenaway¹, Mike Thelwall² and Ying Ding³

¹ *s.greenaway@wlv.ac.uk*

School of Computing and Information Technology, University of Wolverhampton, Wolverhampton (UK)

² *M.Thelwall@wlv.ac.uk*

School of Computing and Information Technology, University of Wolverhampton, Wolverhampton (UK)

³ *dingying@indiana.edu*

School of Library and Information Science, Indiana University, Indiana (USA)

Abstract

A problem exists of how to categorise the abundance of user generated content being uploaded to social sites. One method of categorisation being applied is tagging, user generated keywords that are assigned to the content. This research presents a study into the tagging practice of YouTube users. A classification scheme was applied to a dataset of 768 tags, assigning the tags to different categories of tag type. Analysis reveals how useful the tagging method on YouTube is at improving the categorisation of user generated video content in contrast to collaborative tagging systems.

Introduction

One of the key attributes of Web 2.0 sites, in conjunction with social networks, is tagging; a useful tool for labelling online resources like web pages, audio, images and video. This article is primarily concerned with online video. Video sharing websites such as YouTube, GoogleVideo and MySpace.tv provide User Generated Content (UGC) to mass audiences. The diversity of user generated video creates difficulties for categorisation and findability. Current methods of searching for internet video using existing keyword search techniques are inadequate because of the lack of meta data available for videos. Titles, descriptions, social information and minimal existing tag data are insufficient to accurately describe the content of video (Yang et al. 2007).

At present tagging on YouTube is not collaborative, with only the owner of the video being able to tag. If collaborative tagging was introduced, any user could tag any video, increasing the number of tags for a video and potentially improving video search. Geisler and Burns (2007) found that YouTube tags added additional description of the video content that was not found in other text on the page. Halvey and Keane (2007) found that few users interact with the social element of the site i.e. join groups, upload videos, make friends, favourite videos or comment. If few users upload content to many viewers, only the tags of a few users are being used as additional textual data for the videos. Therefore, whilst there is potential for more tags to be entered and for rich folksonomies to be created, there remains the problem that if only a few users interact with the social elements of YouTube, how can users be encouraged to tag?

This research presents an analysis of tagging behaviour on YouTube, through a classification of the user generated tags assigned to a random selection of 100 YouTube videos. Tags were classified into various categories of tag type, using a custom classification scheme. The research investigates whether theories of structure, motivation and tag type applied to collaborative tagging systems (Golder & Huberman, 2006, Marlow et al. 2006, Angus et al. 2008) are evident in YouTube tag data. It is a preliminary study into understanding how useful the tags entered by the uploader of the video are at describing the content to other YouTube users and if the absence of collaborative tagging has an impact on the types of tag and the cognitive level of the vocabulary.

Background

Golder and Huberman (2006) claim that the main problem with tagging stems from its free-form nature. The absence of any controlled vocabulary means that tags have a multitude of different spellings, plurals, terminology, opinions, descriptions, dialects and languages. Croft and Cruse (2004) argue that words can be categorised based on their level of specificity, or cognitive level. When applied to tags, there are three cognitive levels superordinate, basic and subordinate. Basic level tags have the least cognitive cost to the user – that is they are thought of more quickly. They are more likely to have a high frequency as there is a greater probability for agreement of terms than subordinate level tags (Golder & Huberman, 2006).

Collaborative tagging of images on the Flickr website has provoked research into tagging behaviour, types of tag and semantic relations, (Aurnhammer et al., 2006; Marlow et al., 2006; Rafferty & Hilderly, 2007; Ames & Naaman, 2007, Angus et al., 2008). The research has revealed the quantity and diversity of tags entered by both resource owner and other users. Research into tagging on YouTube is not as extensive as that of Flickr. Research centres on quantitative analysis of YouTube tags (Geisler & Burns, 2007; Halvey and Keane, 2007; Paolillo, 2008) rather than focussing on the vocabulary of the tags.

Ding et al. (submitted) highlight a problem with analysing tags in YouTube; because only the user uploading the video can tag, there is no indication of the collaborative opinion of viewers of the video. YouTube tags can only indicate trends in the type of content being uploaded to the site, but cannot offer insight into the type of content users prefer watching. The authors note that using tag frequency to identify community interest is not possible in YouTube.

Methods

Data Collection

The dataset of Ding et al. (2008) was used for this study. The data was originally collected as follows: In September 2007 a crawl of YouTube was conducted to obtain a dataset of video URLs and tagging data. The crawler started from the main page at <http://youtube.com> and visited every available video page (links starting with <http://www.youtube.com/watch?v>). On each video page it collected tagging data and visited the links pointing to other video pages. YouTube does not provide related tag data. In order to avoid visiting the same page more than once, the query parts of links were ignored.

The original dataset contained 43,641 tags. The majority of foreign words or characters in particular, Chinese/Japanese characters that had not converted correctly into the text file were manually removed; 1,461 entries were removed leaving a dataset of 42,180 tags. A random selection of 100 videos and their assigned tags were then extracted from the dataset using a custom script. This created a dataset of 768 tags for Classification.

Classification Scheme

Angus et al. (2008) developed a classification scheme based on possible image categories in Flickr. For the purposes of this research, the classification scheme was modified to be more suited to a classification of YouTube Tags. The distinction between social and personal motivation was removed, with categories in A and B being tags generally descriptive of the content and categories in C being of use only

to specific users or groups within the YouTube community. Rather than miscellaneous categories as defined by Angus et al. (2008), categories in D are tags which are either irrelevant, or seen as not useful in terms of describing or indentifying the video in search or tag browsing. Alongside restructuring the classification scheme, five new categories were added.

Tags were classified whilst watching the respective video, by a single classifier. Some videos were no longer available, and so the tags assigned to these videos were classified into the D5 (unable to determine relationship) category.

Findings

A large number of tags referred to people, this is not depicted by the B1b (people/animals/objects) result of 9.5% as the majority of these tags were classified into the D2 (Multi-words) category. The largest percentage of tags, 23.3%, were placed into the D2 category. Some of the tags classified in this category resulted from complete sentences being placed in the tag field, either as a description of the content or the title. Considering this tagging practice by users, a surprisingly low result of 3.3% was recorded for the D7 (Conjunctions and Prepositions) category. It had been expected that a higher percentage of these tags would be found in relation to the other categories, due to the finding in Ding et al. (submitted) that *'the'* is the most frequently assigned tag for the years 2006 and 2007 and fourth in 2005.

Table 1 – Total number of tags and corresponding percentage of all tags, for each classification category.

Classification Category		No of tags	Percentage of all tags
A	Generic relationship between tag and video content		
A1	Tag generically identifies what video is 'of'	85	11.1%
A2	Tag identifies video Category/Genre	42	5.5%
B	Specific relationship between tag and video content		
B1a	Tag specifically identifies what video is 'of' (place names/events)	66	8.6%
B1b	Tag specifically identifies what video is 'of' (people/animals/objects)	79	9.5%
B2	Tag identifies what video is 'about'	67	8.7%
B3	Tag identifies opinion expression	51	6.6%
C	Tag only useful to a minority of users, specific individual or group		
C1	Refining tag	45	5.9%
C2	Self-reference tag	5	0.7%
C3	Tag which explicitly denotes ownership of video	8	1%
D	Irrelevant/Non Useful Tags		
D1	Compound tag (truncating or compounding words to form one tag)	3	0.4%
D2	Multi-word tags (individual words in these)	179	23.3%
D3	Attention attracting tags	3	0.4%
D4	Misspelling	4	0.5%
D5	Unable to determine relationship	39	5.1%
D6	Foreign word/character	67	8.7%
D7	Conjunctions and prepositions	25	3.3%

Category A1 (what the video is of) and A2 (category/genre) will contain mostly basic level tags that describe the content at its most general. 11.1% of all tags were classified A1 and was the second highest category. Surprisingly, A2 contained only 5.5% of tags, suggesting that YouTube taggers describe the video content more than they use tagging to categorise the video, using the pre-assigned YouTube categories only. This finding is emphasised by the high percentage of Category B tags, that more specifically describe the video content and may require some specialist knowledge. B1b (9.5%), B2 (what the video is about) contained 8.7% of tags, B1a (places/events) contained 8.6% and B3 (opinion expression) 6.6% of all tags. An indication that YouTube taggers use more specific level vocabulary over basic level generalised terms is that 5.9% of tags were classified as C1 (refining tag) tags. The tendency of YouTube taggers to use more subordinate level, descriptive tags could explain the low percentage, 0.4% of category D3, attention attracting tags. It would be expected that these tags would be of basic level vocabulary, maximising the probability of agreement on terms, with tags being words that are perceived to be regularly searched for, or relate to popular categories or videos. To accurately assess the specificity of the tag vocabulary, tag frequency and co-occurrence metrics can be analysed (Golder & Huberman, 2005; Cattutto, 2007). This is not possible with this data sample as only 6.6% of tags occur more than once.

Discussion

Collaborative tagging allows for the taggers in the system to classify and categorise the content in the system using language useful to the community. In YouTube this doesn't exist, as only the owner of the video can tag and they may not use language or a style of tagging that is useful to the community. Without collaborative tagging there is no agreement between taggers that tags are good, useful and relevant to the content and as a result there is no reuse of tags by which to measure tag relevance.

More multi-word tags were identified than compound tags, this can be seen as a positive tagging behaviour of YouTube users. Multi-word tags are more useful in keyword search than compound tags as users are unlikely to enter the compounded word as a search term. Multi-word tags can also be useful to create long-description meta data for videos that can improve indexing of videos. However, there is a usability problem of how to accept and handle multi-word tags in a tagging system.

Conclusion

The results suggest that YouTube users use tagging as an extension of the description and title fields. Tags do not appear to be used to further categorise a video, with users apparently relying on the categorisation structure of the YouTube system for this purpose. This is surprising since Flickr tags seem to be frequently useful for this purpose (e.g., Angus et al., 2008) and suggests that YouTube video posters are less aware of the need to publicise their work through tags. The classification found that YouTube taggers used a relatively specific vocabulary to describe their videos, for instance, tagging the species of dinosaur, rather than just tagging dinosaur; or tagging the make and model of motorbike, as opposed to just entering the motorbike tag. Whilst these tags may be useful at finding less popular videos through keyword search, in theory, searchers are unlikely to use more specific vocabulary for keyword terms, so the tags may well be relevant to only a few users rather than the majority (Furnas et al., 1987; Golder & Huberman, 2006). It may not be the case that the syntax used is too specific for the majority of users, but rather that without the collective vocabulary provided by collaborative tagging it is impossible to accurately assess the specificity of the tags or the level of agreement of terms achievable. The lack of agreement between YouTube tags makes the clustering of videos for related content impossible, impacting on their potential for categorising user generated videos.

Through analysis and classification of collaborative tagging data it is possible to evaluate the collective intelligence of the community, to assess the social impact of a resource or user, to discover community interest, trends, popularity and social connections. The method of tagging implemented in YouTube does not allow for such evaluations, and it is not clear why this is the case. With the introduction of a collaborative tagging system it would be possible to assess the popularity of the videos through analysis of the amount of tags entered per video, the type of tag entered, language used and opinions expressed. Trends in viewing habits could be uncovered, which could improve the recommendation of videos. Recommendation systems could be developed based on shared user interest and co-occurrence of tags. The tags themselves could provide a method for categorising the increasing amount of user generated content, either for retrieval, for curating collections, or for preservation of content.

References

- Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, April 28 - May 03, 2007)
- Angus, E., Thelwall, M. & Stuart, D. (2008) General patterns of tag usage among university groups in Flickr. *Online Information review*, 32(1), 89-101
- Aurnhammer, M., Hanappe, P. & Steels, L. (2006) Integrating collaborative tagging and emergent semantics for image retrieval. *Proc. of the Collaborative Web Tagging Workshop (WWW '06)*.
- Cattuto, C., Loreto, V. & Pietronero, L. (2007) Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences (PNAS)*, **104**(5), (pp. 1461-1464).
- Croft, William A. & D.A. Cruse (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- Ding, Y. Toma, I. Kang, S. Zhang, Z. and Fried, M. (2008). [Mediating and Analyzing Social Data](#). *Proceedings of The 7th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE 2008)*, Lecture Notes in Computer Sciences, Nov 11-13, 2008, Monterrey, Mexico, Springer-Verlag.
- Ding, Y., Jacob, E.K., Zhang, Z., Foo, S., George, N.L., Guo, L. & Yan, E. (submitted) Adding Semantics to Social Tags.
- Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), (pp. 964-971).
- Geisler, G. & Burns, S. (2007) Tagging video: conventions and strategies of the YouTube community. *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, (pp. 480-480)
- Golder, S. & Huberman, B. (2006) Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198--208.

- Halvey, M. J. & Keane, M. T. (2007). Analysis of online video search and sharing. In Proceedings of the Eighteenth Conference on Hypertext and Hypermedia (Manchester, UK, September 10 - 12, 2007). HT '07. ACM, New York, NY, 217-226
- Marlow, C., Naaman, M., Boyd, D. & Davis, M. (2006) HT06, tagging paper, taxonomy, Flickr, academic article, to read. HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, (pp. 31-40).
- Paolillo, J. C. (2008). Structure and Network in the YouTube Core. In Proceedings of the Proceedings of the 41st Annual Hawaii international Conference on System Sciences (January 07 - 10, 2008). HICSS. IEEE Computer Society, Washington, DC, 156.
- Rafferty, P. & Hilderly, R. (2007) Flickr and Democratic Indexing: Dialogic Approaches to Indexing. *Aslib Proceedings* **59**(4) (pp.397-410).
- Yang, L., Liu, J., Yang, X., & Hua, X. 2007. Multi-modality web video categorization. In Proceedings of the international Workshop on Workshop on Multimedia information Retrieval (Augsburg, Bavaria, Germany, September 24 - 29, 2007).

Appendix B

Poster presented at SNIC 2009

The Broad Side Of Video Tagging – A Classification Of Tags From YouTube And Viddler

Overview

This research presents a classification study of tags collected from both Viddler and YouTube videos. The aim of the research was to discover if the language of the tags or tag type is different depending on whether the tags were entered into a broad (Viddler) or a narrow (YouTube) tagging system. In addition the research aims to analyse whether the tag type is affected by the category of the video.

Research Questions

1. Is there evidence of collaborative tagging on viddler?
2. Does collaborative tagging have an impact on tag type?
3. Does the category of video have an impact on tag type? Do people tag Entertainment videos differently to Informative videos?

Category Type	YouTube	Viddler
Entertainment	Comedy	Comedy
	Entertainment	Entertainment
	Gaming	Game/Games
	Music	Music
Informative	Technology	Technology
	Sport	Sports
	Travel	Travel
	News	News

Table 1 Video Categories

Methods

Eight YouTube categories were chosen and their respective tag found in Viddler. Viddler does not have pre-defined categories to choose from so the corresponding tag had to be used (see Table 1).

The YouTube API and Viddler API were used to capture a dataset of unique videos and associated textual data (title, user, views and tags). For YouTube, 300 of the 'most recent' videos in each category were retrieved each day over a 5 day period. For Viddler the feed was ordered by 'most recent', all videos were retrieved in each of the 8 categories on one day. The dataset was cleaned, removing videos that had been deleted from the system, or contained 100% non-English word/character tags. This left a YouTube dataset of 10,870 unique videos and a Viddler dataset of 46,573 unique videos.

Table 2 - Classification Scheme

A random selection of 1 tag each from 100 videos were extracted from each category, from both datasets. A custom classification scheme was used to classify the tags for tag type and language parameters. Analysis of the classification will discuss the specificity of the vocabulary of the tags and whether this is affected by first the tagging system and second, the video's category.

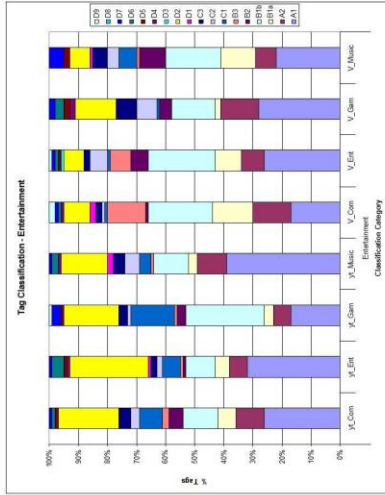
Pre-Representative	Objective	Subjctive
Iconographic	Basic	Specific
	1. Tag generally identifies what video is 'of'	1. Tag specifically identifies what video is 'of' (place)
	2. Tag identifies video Category/Genre	2. Tag identifies what video is 'about'
Interpretive	1. Tag identifies video Category/Genre	1. Tag identifies Opinion Expression
	2. Tag identifies what video is 'of' (place)	2. Tag only useful to a minority of users, specific individual
	3. Tag identifies what video is 'about'	3. Self-reference tag
None	1. Tag identifies what video is 'of' (place)	1. Video Year Tag
	2. Tag identifies video Category/Genre	2. Tag which explicitly denotes ownership of video
	3. Tag identifies what video is 'about'	3. Video Year Tag
	4. Tag identifies what video is 'of' (place)	4. Attraction Attributing Tag
	5. Tag identifies video Category/Genre	5. Misspelling
	6. Tag identifies what video is 'about'	6. Foreign word/character
	7. Tag identifies what video is 'of' (place)	7. Conjunctions and prepositions
	8. Tag identifies video Category/Genre	8. Unrelated tag
	9. Tag identifies what video is 'about'	9. URL

Findings

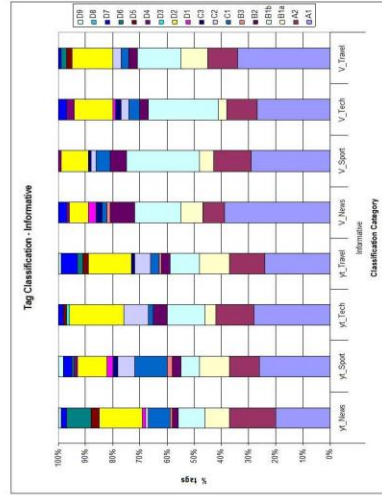
- Entertainment has 68% more Opinion Expression (B3) tags than Informative. Viddler has 42% more B3 tags than YouTube.
- Informative has 28% more Conjunctions and Prepositions (D7) than Entertainment.
- Entertainment has 56% more tags that denote ownership (C3) than Informative.
- Entertainment has 19% more subjective (B2+B3) tags than Informative.
- Viddler has 34% more subjective (B2+B3) tags than YouTube.
- YouTube has 28% more Multitword tags (D2) than Viddler. Marginally (4%) more in Entertainment than Informative.
- Viddler has 14% more tags that denote ownership (C3) than YouTube.
- YouTube has 46% more Foreign Word tags (D6) than Viddler.
- Viddler has 24% more B1b tags than YouTube. Marginally more (4%) in Entertainment than Informative.
- Viddler has 30% more Interpretive B2 tags than YouTube and there are 7% more in Informative than Entertainment.
- YouTube has 44% more Refining (C1) tags than Viddler.

Conclusion

- There is no clear evidence that collaborative tagging is used on Viddler as it is not possible to find out who assigned the tag.
- The classification findings suggest that Viddler users tag with more specific vocabulary at a more subjective, rather than objective level. This may imply that some videos are collaboratively tagged.
- The study found that people are more likely to tag Entertainment videos subjectively, with Informative videos being tagged at a more basic level.
- Further work will analyse the data further to find any correlation between no. views and no. of tags, to find evidence of collaborative tagging. A statistical analysis of the classification results will be conducted to try and empirically answer the research questions.



Graph 1 – Results of tag classification for Entertainment videos.



Graph 2 – Results of tag classification for Informative videos.

Appendix C

Usability Questionnaire SUS evaluation.

1. I think that I would like to use this website frequently:
2. I found the website unnecessarily complex:
3. I thought the website was easy to use:
4. I think that I would need the support of a technical person to be able to use this website:
5. I found the various functions in this website were well integrated:
6. I thought there was too much inconsistency in this website:
7. I would imagine that most people would learn to use this website very quickly:
8. I found the website very cumbersome to use:
9. I felt very confident using the website:
10. I needed to learn a lot of things before I could get going with this website:

Desirability Questionnaire

MEASURING DESIRABILITY

Please select all words that best describe your experience using VideoTag:

- | | | | | |
|--|---------------------------------------|---|---------------------------------------|---|
| <input type="checkbox"/> Accessible | <input type="checkbox"/> Creative | <input type="checkbox"/> Fast | <input type="checkbox"/> Meaningful | <input type="checkbox"/> Slow |
| <input type="checkbox"/> Advanced | <input type="checkbox"/> Customizable | <input type="checkbox"/> Flexible | <input type="checkbox"/> Motivating | <input type="checkbox"/> Sophisticated |
| <input type="checkbox"/> Annoying | <input type="checkbox"/> Cutting edge | <input type="checkbox"/> Fragile | <input type="checkbox"/> Not Secure | <input type="checkbox"/> Stable |
| <input type="checkbox"/> Appealing | <input type="checkbox"/> Dated | <input type="checkbox"/> Fresh | <input type="checkbox"/> Not Valuable | <input type="checkbox"/> Sterile |
| <input type="checkbox"/> Approachable | <input type="checkbox"/> Desirable | <input type="checkbox"/> Friendly | <input type="checkbox"/> Novel | <input type="checkbox"/> Stimulating |
| <input type="checkbox"/> Attractive | <input type="checkbox"/> Difficult | <input type="checkbox"/> Frustrating | <input type="checkbox"/> Old | <input type="checkbox"/> Straight Forward |
| <input type="checkbox"/> Boring | <input type="checkbox"/> Disconnected | <input type="checkbox"/> Fun | <input type="checkbox"/> Optimistic | <input type="checkbox"/> Stressful |
| <input type="checkbox"/> Business-like | <input type="checkbox"/> Disruptive | <input type="checkbox"/> Gets in the way | <input type="checkbox"/> Ordinary | <input type="checkbox"/> Time-consuming |
| <input type="checkbox"/> Busy | <input type="checkbox"/> Distracting | <input type="checkbox"/> Hard to Use | <input type="checkbox"/> Organized | <input type="checkbox"/> Time-Saving |
| <input type="checkbox"/> Calm | <input type="checkbox"/> Dull | <input type="checkbox"/> Helpful | <input type="checkbox"/> Overbearing | <input type="checkbox"/> Too Technical |
| <input type="checkbox"/> Clean | <input type="checkbox"/> Easy to use | <input type="checkbox"/> High quality | <input type="checkbox"/> Overwhelming | <input type="checkbox"/> Trustworthy |
| <input type="checkbox"/> Clear | <input type="checkbox"/> Effective | <input type="checkbox"/> Impersonal | <input type="checkbox"/> Patronizing | <input type="checkbox"/> Unapproachable |
| <input type="checkbox"/> Collaborative | <input type="checkbox"/> Efficient | <input type="checkbox"/> Impressive | <input type="checkbox"/> Personal | <input type="checkbox"/> Unattractive |
| <input type="checkbox"/> Comfortable | <input type="checkbox"/> Effortless | <input type="checkbox"/> Incomprehensible | <input type="checkbox"/> Poor quality | <input type="checkbox"/> Uncontrollable |
| <input type="checkbox"/> Compatible | <input type="checkbox"/> Empowering | <input type="checkbox"/> Inconsistent | <input type="checkbox"/> Powerful | <input type="checkbox"/> Unconventional |
| <input type="checkbox"/> Compelling | <input type="checkbox"/> Energetic | <input type="checkbox"/> Ineffective | <input type="checkbox"/> Predictable | <input type="checkbox"/> Understandable |
| <input type="checkbox"/> Complex | <input type="checkbox"/> Engaging | <input type="checkbox"/> Innovative | <input type="checkbox"/> Professional | <input type="checkbox"/> Undesirable |
| <input type="checkbox"/> Comprehensive | <input type="checkbox"/> Entertaining | <input type="checkbox"/> Inspiring | <input type="checkbox"/> Relevant | <input type="checkbox"/> Unpredictable |
| <input type="checkbox"/> Confident | <input type="checkbox"/> Enthusiastic | <input type="checkbox"/> Integrated | <input type="checkbox"/> Reliable | <input type="checkbox"/> Unrefined |
| <input type="checkbox"/> Confusing | <input type="checkbox"/> Essential | <input type="checkbox"/> Intimidating | <input type="checkbox"/> Responsive | <input type="checkbox"/> Usable |
| <input type="checkbox"/> Connected | <input type="checkbox"/> Exceptional | <input type="checkbox"/> Intuitive | <input type="checkbox"/> Rigid | <input type="checkbox"/> Useful |
| <input type="checkbox"/> Consistent | <input type="checkbox"/> Exciting | <input type="checkbox"/> Inviting | <input type="checkbox"/> Satisfying | <input type="checkbox"/> Valuable |
| <input type="checkbox"/> Controllable | <input type="checkbox"/> Expected | <input type="checkbox"/> Irrelevant | <input type="checkbox"/> Secure | |
| <input type="checkbox"/> Convenient | <input type="checkbox"/> Familiar | <input type="checkbox"/> Low Maintenance | <input type="checkbox"/> Simplistic | |

Submit

Playability Questionnaire

1. I felt that VideoTag was sufficiently challenging for me.
2. I felt that the level of challenge increased as the game progressed.
3. VideoTag is able to challenge people with different skill levels.
4. I found VideoTag challenging even after playing many rounds.
5. I found that the game interface is simple and well-designed.
6. I felt bored when I was playing VideoTag.
7. I found that VideoTag was difficult and stressful.
8. I was able to stay focused on the game tasks.
9. VideoTag was intellectually stimulating.
10. I was motivated by the given time-limit and/or scoring system of the game to continue playing.
11. The actions I took in VideoTag could impact my score.
12. The design of VideoTag prevents serious errors from occurring.
13. I was able to recover from errors that I made without affecting the operation of the game.
14. I could learn quickly how to play VideoTag.
15. I could play Video Tag without reading the instructions.
16. I found that learning to play VideoTag was part of the fun.
17. Help was available when I was faced with difficulties in the game.
18. VideoTag is a useful tool for creating new keywords for videos.
19. VideoTag encourages me to create new keywords for videos.
20. VideoTag is worth playing.
21. I enjoy playing VideoTag.
22. I will continue to play VideoTag if it is available.
23. I preferred playing Top Tag over Golden Tag
24. I understood the difference in game play between Golden Tag and Top Tag.
25. I preferred to tag videos in Simply Tag rather than play a game.

Appendix D

VideoTag Tag Classification Instructions

The classification scheme detailed in Table 7-1 has been created to find out how tags entered for videos describe the video content. The aim is to create extra textual descriptions of the videos using tags. Useful tags could match keywords entered by users searching for the video. It would be helpful to familiarise yourself with the scheme before reading further.

The scheme has three categories of tags, tags that describe the content of the video at a basic level, tags that describe it at a specific level and tags that are not useful or irrelevant to the content. Tags that describe the video do so at either an objective or subjective level, where objective tags identify objects in the video, subjective tags interpret the video content. Objects in this sense can be inanimate objects in the traditional sense but also people, animals, places, events, actions or speech. Anything that appears in the video that is transcribed in a tag is useful at describing the video and is therefore relevant. Tags that are not useful have a social or organisational purpose rather than describing the content e.g. myvideo, or contain spelling mistakes for instance. Tags containing two or more words are classed as multi-word tags and classified as either B1b or D2 accordingly (see the example below). The whole phrase is considered as one tag and not the individual words. Table 1 gives full descriptions of each tag classification category; please refer to the scheme and these notes to be certain of each classification.

The tags to be classified can be found in the attached excel spreadsheet. Alongside each tag is the YouTube video ID (use to watch the video), the video title (use for categorising basic tags), the YouTube or VideoTag category of the video (useful for categorising A2 tags) and the user who uploaded the video or the video owner (useful for C tags). This information provides additional textual data for the video and is required to classify the tags. Please enter the classification label e.g. A1, A2 etc in the column labelled 'classification'.

For each tag the video needs to be watched for up to one minute to see if it identifies an object in the video or interprets the whole video. Some videos may need to be watched for the full duration. The tag needs comparing to existing textual data (available in the excel spreadsheet) if it is not easily categorised and to confirm its specificity. If it can be classified by the accompanying text the video does not need to be watched and the tag is recorded as basic. If multiple tags have the same YouTube ID, the video need only be watched once as long as the tags can be classified accurately.

Use the following URL to watch the video, changing the video ID (highlighted in red) to the YouTube ID next to the tag in the excel spreadsheet.

<https://www.youtube.com/watch?v=qybpwgKwIJ4>

Before assigning a tag to a D5 'no relevance' category please watch the video in full to make sure it has no relevance. Also if you are unsure of a category please watch the video again. A tag is relevant if it describes something that appears in the video,

the audio or interprets the video content. How specifically it describes the content is the most difficult categorisation. If a tag belongs to two categories please choose the category you think fits best using the guide below:

A video shows a man and a boy playing football in the park. A dog runs over and steals the ball and bursts it. The boy cries. The title of the video is "Victory Rover – Kick about in West Park" it was uploaded by MrP3456.

The following tags are entered:

'man', 'boy', 'dog', 'Peter', 'Harry', 'football', 'Rover', 'ball', 'West Park', 'park', 'green', 'trees', 'kick', 'red t-shirt', 'blonde', 'spaniel', 'play', 'bite', 'chasing', 'burst', 'crying', 'MrP', 'disappointment', 'funny', 'PeterSmith'.

A1 tags would be 'man', 'boy', 'dog', 'ball', 'park', 'green', 'trees', 'football'.

B1b tags would be 'Peter', 'Harry', 'blonde', 'spaniel'.

B1c tags would be 'bite', 'burst', 'crying', 'chasing', 'play'.

B1d would be 'red t-shirt'

'West Park' would be D2 not B1a because it a multi-word tag and not B1d because it appears in the video title, it has not taken specific knowledge to think of the tag. To be assigned to D2 the entire phrase must appear in the title. For instance if the tag was 'family fun in west park' this would be a B1d because the whole phrase adds a textual description, the tag 'west park' on its own just repeats existing available text. The same is true for 'Rover' which would be A1 not B1b and 'kick' which would be A1 not B1c.

'disappointment' would be B2 as the tag interprets the video content describing the boys upset.

'funny' would be B3 because it is the opinion of the tagger.

'Mr P' would be C3 as it denotes ownership of the video.

'PeterSmith', is a compound tag. Even though it specifically describes a person in the video it is not a B1b and whilst it could denote ownership (C3), its primary function is a compound tag and is therefore classified as a D1.