The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Promotiecommissie:
prof.dr. A. Th. Schreiber (promotor)
prof.dr. L.M. Aroyo (promotor)
dr. J.R. van Ossenbruggen (copromotor)
prof.dr. F.A.H van Harmelen
prof.dr.ir G.J.P.M. Houben
dr. R. Ordelman
prof.dr. E. Simperl
prof.dr. C.G.M. Snoek
VRIJE UNIVERSITEIT

Serious Games in Audio-Visual Collections

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètwetenschappen
op dinsdag 20 november 2018 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Risto Gligorov

geboren te Valandovo, Macedonië
promotoren: prof.dr. A.Th. Schreiber
          prof.dr. L.M. Aroyo
copromotor: dr. J.R. van Ossenbruggen
# Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Context and Research Problems</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Audio-visual Collections in the Digital Era</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Project context</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3 Research questions</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Related Work</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 Games With a Purpose</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 Evaluation of End-user Tags in Professional Environment</td>
<td>5</td>
</tr>
<tr>
<td>1.2.3 User annotations for search</td>
<td>5</td>
</tr>
<tr>
<td>1.2.4 Quality and Refinement of Annotations</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Approach</td>
<td>7</td>
</tr>
<tr>
<td>1.3.1 Waisda? User Tags vs. Professional Annotations</td>
<td>7</td>
</tr>
<tr>
<td>1.3.2 Investigating the Added-value of User Tags for Fragment Search</td>
<td>8</td>
</tr>
<tr>
<td>1.3.3 User Tag Quality Metrics</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Structure of the Thesis</td>
<td>9</td>
</tr>
<tr>
<td>1.5 Publications</td>
<td>10</td>
</tr>
<tr>
<td>1.6 Contributions</td>
<td>11</td>
</tr>
<tr>
<td>2 Game Tags in Audio-Visual Collections</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Related work</td>
<td>15</td>
</tr>
<tr>
<td>2.2.1 Games with a purpose</td>
<td>15</td>
</tr>
<tr>
<td>2.2.2 Evaluation of end-user tags</td>
<td>16</td>
</tr>
<tr>
<td>2.2.3 Classification of user descriptions</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Approach</td>
<td>17</td>
</tr>
<tr>
<td>2.4 Materials</td>
<td>18</td>
</tr>
<tr>
<td>2.4.1 Waisda? data snapshot</td>
<td>18</td>
</tr>
<tr>
<td>2.4.2 Domain and lexical vocabularies</td>
<td>19</td>
</tr>
</tbody>
</table>
2.4.3 Videos ................................................. 20
2.4.4 Subtitles .............................................. 20
2.5 Experiments ............................................. 21
  2.5.1 Matching tags to vocabularies ..................... 21
  2.5.2 Tags in subtitles .................................. 24
  2.5.3 Tag classification ................................ 25
2.6 Conclusions and Discussion ........................... 28

3 Game Tags for Visual Video Search ..................... 31
  3.1 Introduction ........................................... 32
  3.2 Related Work ......................................... 33
  3.3 Approach .............................................. 34
  3.4 Datasets and Resources ............................... 35
    3.4.1 The MBH Video and Metadata Collection ........ 35
    3.4.2 Evaluation Dataset ................................ 36
  3.5 Experiments .......................................... 41
    3.5.1 Experiment 1 ...................................... 41
    3.5.2 Experiment 2 ...................................... 42
  3.6 Results ................................................ 42
    3.6.1 Experiment 1 ...................................... 42
    3.6.2 Experiment 2 ...................................... 45
  3.7 Conclusions .......................................... 46

4 Game Tags for Topical Video Search .................... 49
  4.1 Introduction .......................................... 50
  4.2 Related Work ......................................... 52
  4.3 Approach .............................................. 56
  4.4 Experimental Data .................................... 58
    4.4.1 The MBH Video and Metadata Collection ........ 58
    4.4.2 Evaluation Dataset ................................ 59
  4.5 Evaluation of Game Tags for Topical Video Search .... 62
    4.5.1 Game tags vs. catalog data ....................... 62
    4.5.2 Results ........................................... 63
  4.6 Filtering Non-topical Game Tags ...................... 67
    4.6.1 Tag Filters ........................................ 68
    4.6.2 Search Engines .................................... 72
    4.6.3 Results ........................................... 72
  4.7 Conclusion and Discussion ........................... 76
5 **Waisda?: Open-source Video Labelling Game**  
5.1 Introduction ........................................ 79  
5.2 Gameplay and User Interface .................. 81  
  5.2.1 Homepage .................................. 81  
  5.2.2 Game ....................................... 82  
  5.2.3 Game Recap ................................ 83  
5.3 Implementation .................................. 83  
  5.3.1 Backend .................................. 85  
  5.3.2 Frontend .................................. 86  
5.4 Waisda? Deployments ............................. 87  
  5.4.1 Europeana .................................. 87  
  5.4.2 Spotvogel .................................. 88  
  5.4.3 European Film Gateway Experiment ....... 89  
5.5 Thesis Contributions .............................. 89  
  5.5.1 Software Development ....................... 89  
  5.5.2 Game Design ................................ 93  
5.6 Conclusion ........................................ 94

6 **Conclusions and Discussion** ................. 95  
6.1 Main Contributions ............................... 95  
6.2 Research Questions Revisited .................. 97  
  6.2.1 What is the relationship between the game tags and  
  professional annotations in terms of vocabulary and  
  what they describe? ................................ 97  
  6.2.2 Can we improve video retrieval with the help of user-generated data? .......... 99  
  6.2.3 Can the quality of user-generated tags for videos be  
  evaluated and improved? ......................... 102  
6.3 Discussion and Further Research .............. 102  
  6.3.1 A broader perspective ....................... 103  
  6.3.2 Future Directions ........................... 107

**Appendices** .......................................... 109  
A Precision/Recall per Query for Visual Search .... 111  
B Query Precision over Time for Visual Search .... 119  
C Query Recall over Time for Visual Search ........ 121  
D Precision/Recall per Query for Topical Search .... 123
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td>130</td>
</tr>
<tr>
<td>Summary</td>
<td>147</td>
</tr>
<tr>
<td>Samenvatting</td>
<td>151</td>
</tr>
<tr>
<td>SIKS Dissertatiereeks</td>
<td>155</td>
</tr>
</tbody>
</table>
Preface

I have always believed that research is inherently a collaborative effort. Ideas are rarely born in a split second of epiphany, but are more often shaped and crystallized incrementally over many, sometimes heated, discussions with peers. My experience in doing the work that led to this thesis has confirmed this belief. This thesis would not have been possible without the contributions and support of many people. First and foremost I would like to thank my supervisors, Guus Schreiber, Lora Aroyo, and Jacco van Ossenbruggen. Guus you offered me enough freedom to find my way and fulfill my potential as a researcher. Your insight in defining the overarching goals of this thesis was instrumental and your guidance during the overall research process was invaluable. I would like to thank Lora for her many excellent ideas and suggestions. Her tireless collaboration with various stakeholders such as the Netherlands Institute for Sound and Vision and several dutch broadcasters greatly contributed to the success of this work. Thank you Jacco for your most valuable contributions and suggestions to both the research and writing process. Many times when I was facing a difficult decision I relied on your judgment which was almost always right.

I would also like to thank Michiel Hildebrand which acted as my forth unofficial supervisor. He joined the Vrije Universiteit Amsterdam as a post-doctoral researched at the end of the first year of my PhD studies. As a recent PhD graduate, he was able to provide a perspective on what it means to be a PhD researcher that resonated with me. This was rather fortunate as I was just at beginning my journey. His suggestions and advice helped to shape many of the results in this work.

Many thanks to Davide Ceolin and Valentina Maccatrozzo Ceolin who were fellow PhD students and are dear friends. We spent many moments together sharing the joy and the occasional torments which are inseparable part of the journey of every PhD student. Your support and friendship are greatly appreciated.

My thanks also to Zarko Aleksovski who was one of my supervisors during my internship at Philips Research. During this period I got acquainted with
the area of Semantic Web and had the pleasure to briefly work with Frank van Harmelen. This experience was the main factor that led me to pursue a PhD in this area. I would also like to thank Jasen Markovski for recommending me for the internship position and many thanks to Dragan Sekulovski for being a great company.

I would like to express special thanks to my dear friend Balthasar Schopman. We worked together for more than two years and his sense of humor and positive outlook on life were a great addition to the working environment. I find his uncompromising attitude towards things that matter to him most admirable.

Special thanks to Ognen Lukas and Daniela Remenska, dear friends with whom I spent many great moments together. They were big and very important part of my life outside of academia.

I would also like to thank all colleagues from the Web and Media and Knowledge Representation and Reasoning groups which I have not mentioned by name here. The discussions and the exchange of ideas during the various meetings and the encounters in the coffee corners were an essential driving force in my work.

Last but not least, I would like to express profound gratitude to my family. To my parents Spomenka and Kole who have sacrificed a great deal to provide me with opportunities and without their endless support and guidance through life I would not be where I am now. To my sister Gonca and her family for always being there for me. To my in-laws Lidija, Emilija and Dragan who have supported me tremendously over the years. And to my beloved Irina and daughter Sofija who are my source of joy and inspiration.
Chapter 1

Introduction

In this chapter we introduce some of the motivating factors and the context of the research work presented in this thesis. We outline the three research questions this thesis is centered around and we describe the methods used to address them. We also present the related work and state the high-level contributions of the thesis.

This chapter is based on the paper entitled *User-generated metadata in audio-visual collections* which was presented at the Doctoral Symposium of the 21st International World Wide Web Conference (WWW2012), held in Lyon, France.

1.1 Context and Research Problems

The central goal of this research work is to investigate the added value of user-generated metadata in professionals environments such as audio-visual archives.

1.1.1 Audio-visual Collections in the Digital Era

Audio-visual (AV) content collections are undergoing a transformation from archives of analogue materials to very large stores of digital data accessible online. Prerequisite for successful information retrieval and collection management is quality metadata associated with collection items. Traditionally, the task of annotating video programs is strictly in the hands of professional cataloguers who adhere to well-established guidelines in the cataloguing process [27]. Usually, the resulting manually-crafted professional annotations are coarse-grained in a sense that they are referring to the entire
program describing the prevalent topics. Fine-grained manual annotation of video fragments, on the other hand, is prohibitive, as the work involved is inevitably tedious, incomplete, and costly [54]. However, a significant portion of user-content requests are targeted at video fragments rather than entire programs; a transaction-log analysis [17] performed in a broadcast archive showed that fragment purchases account for 66% of all purchases. As the program level professional annotations may be inadequate for locating or retrieving of fragments — or any other task that involves fragments rather than entire programs — the need for fine-grained annotations becomes clear. A possible solution to the problem of scarcity of fine-grained video metadata is to harness users’ efforts to amass video fragment-level descriptions.

One of the most common forms of user-generated metadata are tags. With the advent of the Web 2.0, the tagging phenomenon witnessed rapid proliferation in many areas; image collections1, bookmark collections2, user video collections3. Experience has shown that joint efforts of user communities result in massive amounts of tags usually unparalleled — in terms of quantity — by what professionals can provide. Just to put things into perspective, in 200 years of existence the Library of Congress has applied their expert-maintained taxonomy to 20 million books4, whereas, in only four years, Flickr’s users applied their ad-hoc tagging vocabulary to over 25 million photos. This suggests that there is, indeed, a potential in engaging end-users in the video-annotation process. One way of achieving this is through so-called ‘games with a purpose’ [122]. To this end, the Netherlands Institute for Sound and Vision5 (S&V) in cooperation with KRO broadcasting6 launched Waisda? in May 2009, a multi-player video labeling game where players describe streaming video by entering tags and score points based on temporal tag agreements (). The underlying assumption is that tags are probably valid — trustworthily describe the video fragments — if they are entered independently by at least two players within a given timeframe (see figure 1.1). The motivation behind Waisda? is to improve the access to S&V collection [86]. With Waisda? the S&V institute aims to collect metadata in a user vocabulary, as previous research [62] suggests that such metadata can help bridge the gap between the search queries and the indexing vocabulary. In addition, it is expected that the resulting time-related metadata of the content within the video can improve support for finding

---

1http://www.flickr.com
2http://www.delicious.com
3http://www.youtube.com
4http://www.loc.gov/about/reports
5http://portal.beeldengeluid.nl/
6http://www.kro.nl/
1.1. CONTEXT AND RESEARCH PROBLEMS

Figure 1.1: Screenshot of the Waisda? game interface: the central part of the screen is reserved for the video. Immediately below the tag input field and the list of tags entered already are placed. The coloration of the tags is indication of the number of points the tags scored. On the right, there is the list of players currently in the game.

fragments within entire broadcasts [54].

1.1.2 Project context

This research was part of and funded by the PrestoPRIME\(^7\) project which brought together various European audiovisual archives (including S&V), research institutions, and industry partners. The goal of the PrestoPRIME project was to develop practical solutions for the long-term preservation of video and audio broadcasts, and to find ways to increase access by integrating the media archives with European on-line digital libraries. One of the considered ways to increase accessibility to videos was by exploiting user-generated tags collected through the labeling game Waisda?.

1.1.3 Research questions

The overall problem statement for this research is as follows.

\(^7\)http://www.prestoprime.org/
CHAPTER 1. INTRODUCTION

What potential added value do user-generated video annotations have in professional environment?

To successfully integrate user-generated tags into AV collections’ workflows a better understanding of their characteristics, compared to preexisting professionals annotations, is required. In particular, the terminology that users employ when describing videos and the aspects of the video that they usually describe. The first research question, therefore, is

1. What are the relationships between user-generated tags and professional annotations in terms of the vocabulary and what they describe?

Locating a fragment within a video is an important use case in AV collections for which the fine-grained user tags could potentially provide an added value. Thus, the second research question addresses the usefulness aspect of user tags in terms of locating fragments within a video.

2. Can we improve fragment search within video, with the help of user-generated data?

Lastly, successful integrations of user tags also requires that issues around assessing and improving tag quality to be addressed. The third and final research question, therefore, is

3. Can the quality of user-generated tags for videos be evaluated and improved?

1.2 Related Work

In this section we outline some of the related work. It should be noted that this is an incomplete list of the more relevant studies.

1.2.1 Games With a Purpose

Games with a purpose (or GWAPs) are computer games, in which people, as a side effect of playing, perform tasks computers are unable to perform [122]. The first example of a GWAP was the ESP game [121], designed by Luis von Ahn, which harnesses human abilities to label images. The idea to collect metadata through games with a purpose has been applied to video footage in, for example, the Yahoo! video tag game [119], VideoTag, PopVideo and

8http://www.videotag.co.uk/
9http://www.gwap.com/gwap/gamesPreview/popvideo/
1.2. RELATED WORK

Waisda?. Compared to the other video labeling games, Waisda? is unique in the sense that it is initiated by an audiovisual institute (S&V) with the purpose to improve the access to their collection [86].

1.2.2 Evaluation of End-user Tags in Professional Environment

The Steve project [72] was one of first attempts to explore the role of user-generated metadata. In this collaboration of several art museums a collection of artworks was made available to the general public who were asked to tag them. Among other things, the project studied the relationship of the resulting folksonomy to professionally created museum documentation. The results showed that users tag the artworks of art from a perspective different than that of museum documentation: around 86% of tags were not found in museum documentation.

Museum staff also assessed the tags from the steve.museum project on usefulness when used to search for artworks. From the total number of tags, 88.2% were found to be useful. Following the methodology of steve-museum, S&V institute also asked a senior cataloguer to judged a sample of Waisda? tags on their usefulness when searching for videos [86]. The sample consisted of the 20 most frequent and the 20 least frequent tags from two television programs. The cataloguer found the majority of the tags to be useful. She also noted that there seems to exist a difference between professional descriptions and end-user tags. While professionals describe the topical subject of the program, the players in Waisda? generally tag things that can be directly seen or heard in the video. One of the aims of this research is to investigate the characteristics of the tags and what they describe in the video more methodically, and on a larger scale.

1.2.3 User annotations for search

Search based on user-generated metadata, in particular folksonomies, has been studied before. Morrison compared web search performance of folksonomies from social bookmarking web sites against search engines and subject directories [84], showing that search engines had the highest precision and recall rates. Folksonomies, however, performed surprisingly well. In fact, user tags show promise to alleviate the vocabulary mismatch problem for search: bridging the gap between user queries and metadata used for retrieval [34]. Indeed, Geisler and Burns state that YouTube tags provide added value for search, because 66% of them do not appear in other metadata.
Heymann et al. investigated a large-scale sample of forty million bookmarks from the social bookmarking site del.icio.us and found that in 20% of the cases user tags do not occur in the page text, backlink page text, or forward link page text of the pages they annotate. Studies in [11, 43, 94, 135] investigate this phenomenon across multiple domains and multimedia resource types and identify the gaps between the tag space and the querying vocabulary. The common conclusion is that user tags can improve search by bridging the vocabulary gap.

Studies reported in [10, 110] take a more critical stance. They conclude that while overall user tags improve search, not all tags are suitable for retrieval. This hints that a characterization of the quality of the tags is needed to filter out the tags that are not suited for retrieval. This is one of the aspects that we address in this thesis.

Another line of research is exploiting the tripartite structure (Users $\times$ Tags $\times$ Resources) of folksonomies to improve search [53, 7]. Alternatively, the semantics of tags can be grounded in some lexical sources and the grounded tags utilized for improving search. For example, Hildebrand et al. proposed and investigated a semi-automatic process of assigning explicit meaning to user tags for video by linking them to concepts from the Linked Open Data cloud [48].

### 1.2.4 Quality and Refinement of Annotations

There is a substantial body of research into the refinement and quality assessment of annotations of still images. Lee et al. propose a tag refinement technique that aims at differentiating noisy tag assignments from correct tag assignments [73]. Each tag is assigned a probability of being noisy based on the visual similarity of the images and tag co-occurrence statistics. Tags with a probability below a threshold are discarded as noisy.

In [114, 77] neighbour voting schemes for determining the tag relevance are explored. In this approach, a tag is considered more relevant to the image it is ascribed to, also known as the seed image, if the tag is also used to annotate the neighbouring images. The neighbourhood relation is defined in terms of the visual similarity among images. Lee at al. expand the approach by not only considering the visually similar images, but the dissimilar images as well, thus providing negative examples [74]. Kennedy et al. exploits visual similarity among images in a sense that tags ascribed to images by the creators of the image are used as seed annotations and also attached to visually similar images [65]. Zhao et al. propose a data-driven method to automatically determine the relatedness between a tag and the image’s visual content taking into consideration the tag co-occurrence and
the visual similarity among images \cite{139}.

Probabilistic methods that exploit random walk based techniques have also been explored \cite{129, 79, 76}. These methods produce a ranking of the tags according to their relevance with respect to the image with which they are associated. The tag relevance estimations are computed as the stationary or the convergence probabilities of a random walk processes. Notable example of this approach is the PageRank algorithm \cite{26, 67, 63}. Another group of methods exploit background knowledge (such as the lexical database Wordnet and a massive corpus indexed by Google) to perform the refinement of the image annotations \cite{60, 130}. The semantic relations encoded in Wordnet and the semantic similarity quantified by the Google-based measures like the Normalized Google Distance \cite{24} provide contextual evidence for the relationship among the annotations. This evidence is then used as an input for machine learning algorithms which give the final word for the quality of the annotations.

1.3 Approach

In this section we outline the approach to answer the research questions stated above. The specific approach for each of the research questions is described in separate subsection. The order of the subsections is respective to the order in which the research questions are stated in section 1.1.3. At the basis of each of the approaches lies Waisda? which was used as a data-collection tool. In the course of this thesis significant effort went into improving the game design and development of software extension that enables seamless integration of the game into larger workflows. Chapter 5 describes the game in more detail and outlines the contributions of this thesis with respect to the development of the game.

1.3.1 Waisda? User Tags vs. Professional Annotations

To answer the first research question ‘What are the relationships between user-generated tags and professional annotations in terms of the vocabulary and what they describe?’ we perform two studies. Details can be found in \cite{36, 39}.

The first study is a quantitative data analysis of the entire tag collection gathered with Waisda? during the first six months after the game was deployed. In order to estimate the lower bound of the fraction of user tags that are meaningful words, we examine the overlap between them and general
lexical resources and vocabularies. Furthermore, to determine if users and professionals use different vocabularies when describing videos, we investigate the overlap between all user tags and a typical domain thesaurus used by professionals in the cataloging process. The results from the study show that there is, indeed, a terminological gap between users and professionals; while approximately 89% of all user tags are meaningful words only 8% of them were found in the domain thesaurus used by professionals.

In the second study, we take a combined approach. First, we investigate what do users tend to describe more: things heard or things seen on screen. To this end, we perform a study on the overlap between the user tags and the audio signal — subtitles for hearing impaired persons — for a sample of episodes. Second, to get a more comprehensive understanding of the types of tags users usually add, we perform a qualitative study of a sample of user tags obtained through the Waisda? video tagging game. In particular, each tag from the sample is manually analyzed in the light of the video content it describes and categorized in terms of the Panofsky-Shatford classification framework. The results of the study show that user tags predominately describe objects and rarely refer to topics of a scene. This is in sharp contrast with the professional annotations which exclusively target the topic(s) of the entire video and (lot less frequently) of particular scenes.

1.3.2 Investigating the Added-value of User Tags for Fragment Search

Fragment search within video is widely recognized as an application scenario of particular business importance in the AV collections world [17]. However, the existing professional annotations generally refer to the entire video and are not tied to a specific time-point, which decreases their usefulness for fragment retrieval; without temporal data one needs to manually locate the fragment of interest. Waisda? tags, on the other hand, do refer to fragments and are time-based, with time codes that deep link to particular point in the video. Our aim is to investigate their added-value for fragment retrieval. Furthermore, we plan to explore what kind of query tasks (search for events, objects, etc.) benefit from the user tags.

The methodology that we consider is quantitative system evaluation [127]. In order to evaluate effectiveness of information retrieval, this methodology requires collection of “documents” (in our case video fragments), set of queries, and relevance judgments indicating which “documents” in the collection should be returned for each query. We plan to create evaluation dataset from the videos that were part of Waisda? game deployment and tagged
by players. For such a dataset real-life, authentic user tags would already be available. However, set of fragments, relevance judgements, set of queries and query task types need to be defined. The method of achieving this is going to be part of and one of the contributions of the research.

An alternative is to creating an evaluation dataset is using an existing one. There are various video retrieval evaluation initiatives like TRECVID that provide datasets for testing content-based retrieval techniques. The advantages of reusing such a dataset are (i) the set of fragments is defined, (ii) the set of queries and the relevance judgments are defined. The drawbacks are that Waisda? like games have not been run yet on this dataset and setting up such a game and attracting sufficient users for this material could turn out to be hard. Also we are limited to the query task types defined by the dataset. Our position is that the effort required to collect user tags for the material outweighs the advantages offered by this approach.

1.3.3 User Tag Quality Metrics

Our supposition is that any characterization of quality of the user tags is largely determined by how different institutions will use them. In other words, it is immaterial to talk about tag quality without a specific application scenario in mind. Fragment search within video is widely recognized as an application scenario of particular business importance in the AV collections world. For these reasons, our characterization(s) of tag quality will be limited to this scenario. Some of the aspects we plan to investigate in the context of the retrieval scenario are tag frequency and discriminative power of tags, correlation between reputation of players and tag relevance, semantic ambiguity of tags, overlap with transcripts, etc. The exact specifics of this study will be determined by the outcome of the previous studies. Therefore, we plan to make the choice of research methodology and the type of output (e.g. quantitative metrics or set of recommendations) this study will provide after the fragment retrieval study is performed.

1.4 Structure of the Thesis

The thesis is organized as follows. In Chapter 2 we study the characteristics of the tags collected with Waisda?. In particular, we focus on two aspects. First, what is the relationship between the game tags and the annotations created by professional catalogers. Second, we investigate which facets of the video content are typically targeted by the game tags and at what level of

\[10\]http://trecvid.nist.gov/
semantic specificity. In the following two chapters we study the effectiveness of game tags for video retrieval by addressing two prominent search scenario that arise in practice. In Chapter 3 we look into the first scenario which is retrieving videos that feature visual appearances of given objects of interest. In Chapter 4 we focus on the second scenario which is topical search, retrieval of videos that are about a given topic. In Chapter 5 we outline the technical specifications of Waisda? and address two important points. First, how can Waisda? be deployed and used by audio-visual collection owners to tag their collection. Second, how can Waisda? be integrated into larger systems and workflows. Lastly, in Chapter 6 we present the overall conclusions of this thesis. We discuss the implications of this work and suggest directions for future research.

1.5 Publications

Publications on which the chapters of this thesis are based:

• Chapter 1 was published as: Riste Gligorov. User-generated metadata in audio-visual collections. In the Proceedings of the 21st international conference companion on World Wide Web (WWW 2012), Lyon, France, 2012.


• Chapter 3 was published as: Riste Gligorov, Michiel Hildebrand, Jacco Van Ossenbruggen, Lora Aroyo, and Guus Schreiber. An Evaluation of Labelling-Game Data for Video Retrieval. In Advances in Information Retrieval: 35th European Conference on IR Research (ECIR 2013), pages 50-61, Springer, Moscow, Russia, March 2013.


• Chapter 5 is an extension of: Michiel Hildebrand, Maarten Brinkerink, Riste Gligorov, Martijn Van Steenbergen, Johan Huijkm, and Johan Oomen. Waisda?: video labeling game. In the Proceedings of
1.6 Contributions

The main contributions of the thesis are:

- A method to analyze quantitatively the overlap between user and professional terminology exploited in video annotation.

- Qualitative analysis of the aspects of the video content which are described by the game tags i.e. *what do game tags describe in a video and how.*

- A method to design a dataset —including fragments, queries, and relevance judgements — for evaluating the added value of user tags for visual and topical video search.

- Two evaluation datasets —one for visual and one for topical search — which contain fragments, user tags from Waisda? collection, queries, and relevance judgements.

- Quantitative evaluation of the added value of the game tags in terms of visual and topical video search.

- A method for qualitative analysis/classification of the search results in the context of visual search. The aim of the analysis was to discover the types of tags that are generally responsible for false positives.

- Qualitative analysis of the search results in the context of topical search. The aim of the analysis was to discover tagging practices that generally produce relevant topical tags.

- Evaluation of tag-related measures for accessing the quality of the game tags as topical descriptors in the context of topical video search.
• Development of extension module for *Waisda?* which enables integration of the game in AV collection workflows. The module is part of the latest release of *Waisda?* which is available online\textsuperscript{11}.

\textsuperscript{11}The *Waisda?* source code and documentation is available from Github [http://github.com/beeldengeluid/waisda](http://github.com/beeldengeluid/waisda)
Chapter 2

Game Tags in Audio-Visual Collections

In this chapter we study the characteristics of the user tags collected with the video labelling game Waisda?. We address two aspects related to the user tags. First, we investigate to what extent does the terminology employed by the users when tagging videos differ from the terminology used by the professional cataloguers. Second, we examine which facets (who, what, where, and when) of the videos tagged by the users are typically described by the user tags and what is the level of specificness (abstract, general, and specific) of the said tags. The findings in this chapter outline some of the strengths but also the limitations of the user tags. Our analysis shows that there is a terminological gap between the players of Waisda? and the professionals. This makes the user tags a valuable asset as they are from the users and for the users and can improve the access to the videos in the case when the ones accessing are the users themselves. Furthermore, we find that user tags predominately describe instances (objects) in the video and rarely scenes. This fact hints that user tags could be useful for instance search and not so useful for topical search. These matters are addressed in the following chapters.

This chapter was published as “On the Role of User-generated Metadata in Audio Visual Collections” in the Proceedings of the International Conference on Knowledge Capture 2011 [39]. It was co-authored with Michiel Hildebrand, Jacco van Ossenbruggen, Lora Aroyo, and Guus Schreiber.
CHAPTER 2. GAME TAGS IN AUDIO-VISUAL COLLECTIONS

2.1 Introduction

Crowdsourcing has gained attention as a method to collect large numbers of metadata descriptions for media objects [19, 72, 107]. Based on the idea coined by Luis von Ahn [122], a specific type of crowdsourcing has become known as Games With A Purpose (GWAP). Inspired by this idea, the Netherlands Institute for Sound and Vision deployed the video labeling game, Waisda?. Unique for this initiative is that the institute aims [86] to integrate the game into their workflow to complement professional cataloguing and content based retrieval techniques [51]. More specific, with Waisda? they aim to collect metadata in a user vocabulary that describes the content within the video.

We investigate to what extent the aims of Sound and Vision are fulfilled by analyzing the 420,000 user tags collected during the first pilot with Waisda?. To determine the vocabulary used by the crowd, we compare the tags with existing controlled vocabularies. We compare the tags with the professional metadata by matching them to terms of the institutes’ in-house thesaurus. Additionally, by matching the tags to the terms of a Dutch linguistic database, we conclude that a large part of the tags are Dutch words not used by professionals. To determine the type of content that the tags describe we first compare them with the subtitles. Finally, we manually classify the tags from a small number of videos. Using an existing classification model, we show the relation between the content in the video that is described and the type of tags that are used for these descriptions.

The rest of this chapter is structured as follows. Section 2.2 discusses related work. Section 2.3 presents the approach we take in tackling the goals we set forth. Section 2.4 describes the materials we used in our study. Section 2.5 reports on the various experiments we performed on the user tags. Finally, section 2.6 draws conclusions and points to some further directions for research.
2.2 Related work

2.2.1 Games with a purpose

Games with a purpose (or GWAPs) are computer games, in which people, as a side effect of playing, perform tasks computers are unable to perform [122]. The first example of a GWAP was the ESP game [121], designed by Luis von Ahn, which harnesses human abilities to label images. The game randomly pairs up two players with the task to describe images. When both players provide the same label for an image, they score points and proceed to the next image. The labels entered by both users are associated to the image as metadata. In other words, the consensus among users provides a method to ensure the quality and consistency of the labels. Evaluation shows that these labels can be used to retrieve images with high precision and are almost all considered as good descriptions in a manual assessment.

The idea to collect metadata through games with a purpose has been applied to video footage in, for example, the Yahoo! video tag game [119], VideoTag\(^1\), PopVideo\(^2\) and Waisda\(^?\). The gameplay of these video labeling games differs from the ESP game in two ways: (i) multiple users can participate in a single game, and (ii) the users score points when the same tag is entered in a specific time interval. The underlying assumption is that tags are probably valid — trustworthily describe the video fragments — if they are entered independently by at least two players within a given time-frame. From here on we shall refer to tags that are mutually agreed on as verified tags.

Compared to the other video labeling games, Waisda\(^?\) is unique in the sense that it is initiated by an audiovisual institute with the purpose to improve access to their collection [86]. With Waisda\(^?\) the Netherlands Institute for Sound and Vision aims to collect metadata in a user vocabulary, as it is suggested that such metadata can help bridge the gap between the search queries and the indexing vocabulary [62]. In addition, it is expected that the resulting time-related metadata of the content within the video can improve support for finding fragments within entire broadcasts [54]. We investigate to what extent the tags collected in Waisda\(^?\) provide a user vocabulary and analyze what type of content within the video they describe.

\(^1\)http://www.videotag.co.uk/
\(^2\)http://www.gwap.com/gwap/gamesPreview/popvideo/
2.2.2 Evaluation of end-user tags

The steve.museum research [72] was one of first attempts to explore the role of user-generated metadata. In this collaboration of several art museums a collection of artworks was made available to the general public who were asked to tag them. Among other things, the project studied the relationship of the resulting folksonomy to professionally created museum documentation. The results showed that users tag the artworks of art from a perspective different than that of museum documentation: around 86% of tags were not found in museum documentation. We perform a similar study on the collection of *Waisda?* tags by comparing them to in-house thesaurus.

Museum staff also assessed the tags from the steve.museum project on usefulness when used to search for artworks. From the total number of tags, 88.2% were found to be useful. Following the methodology of steve-museum, Netherlands Institute for Sound and Vision also asked a senior cataloguer to judged a sample of *Waisda?* tags on their usefulness when searching for videos [5]. The sample consisted of the 20 most frequent and the 20 least frequent tags from two television programs. The cataloguer found the majority of the tags to be useful. She also noted that there seems to exist a difference between professional descriptions and end-user tags. While professionals describe the topical subject of the program, the players in *Waisda?* generally tag things that can be directly seen or heard in the video. One of the aims of this paper is to investigate the characteristics of the tags and what they describe in the video more methodically, and on a larger scale.

There is substantial body of research work that investigates user tags and folksonomies. For example, in [42, 106] the overall quality of end-user tags is examined and the main strengths (flexibility, simplicity, user perspective, etc.) and potential weaknesses (typos, morphological variation of words, no synonym and no homonym control, etc.) are pin-pointed. Gruber [41] identifies the roles of folksonomies and formal vocabularies and presents use-cases where both can naturally co-exist and cooperate. While many aspects of user tags are well covered in research, little or no attention is paid to the link between tags and the resources they are referring to. In this study we investigate which aspects of the resources (in our case videos) are covered by user tags.

2.2.3 Classification of user descriptions

Various schemes have been developed for classification of user descriptions for visual resources. One of the first is the Panofsky-Shatford model [88, 98] which focuses on the conceptual descriptions. Jaimes and Chang [56] devel-
2.3. APPROACH

oped a classification framework for visual resources (including video) that besides conceptual descriptions also considers perceptual (low-level features) and non-visual descriptions. Hollink at al. [52] combined the previous two schemes and developed a classification framework for user descriptions. As we exploit this framework to classify end-user tags, we explain it in more detail in the following section.

Tag classification framework

The framework distinguishes three top-levels: nonvisual level, perceptual level, and conceptual level. Descriptions at nonvisual level are meant to describe the context of the video but not its content. This is in contrast with descriptions at perceptual and conceptual level which are referring solely to the content of the video. Nonvisual level includes the following classes: creator, title, date, location, carrier type, etc.

Descriptions at perceptual level are derived from low-level audio and visual features of the video. In principle, no domain and no worldly knowledge is required to create descriptions at this level. Perceptual level classes are divided into classes of descriptions that refer to visual features such as color, shape, and texture and classes of descriptions that refer to audio features like volume, pitch, and amplitude.

Descriptions at conceptual level describe the semantic content of the video. To classify tags at this level the Panofsky-Shatford model is used. This model divides conceptual descriptions into three levels: general (generic things in the video), specific (specific things), and abstract (symbolic things). Each of the levels is further broken down into four facets: who, what, where, and when producing the Panofsky-Shatford 3x4 matrix.

In addition, descriptions may be about visual objects or may refer to the entire scene. We take the approach of [56] and define visual objects as entities that can be seen, sometimes differing from the traditional definition of object. Objects like the sky or the ocean would perhaps not be considered objects under the traditional definition, but correspond to our visual objects (as well as the traditional objects like car, house, etc.). Examples of scene descriptions include city, landscape, indoor, outdoor, still life, portrait, etc.

2.3 Approach

We divided our study of the Waisda? data in two parts. In the first part we focus on the user tags, investigating the vocabulary that users employ when describing videos. We analyse the relationship to the vocabularies
used by professional cataloguers and general web users. In the second part
we focus on what the users describe. We analyse which aspects of the video
are described and what type of tags are used for this.

With respect to the first part, we perform the following experiments.
First, in order to estimate the lower bound of the fraction of user tags that
are proper words, we examine the overlap between them and a general lexicon
of the Dutch language. Furthermore, to determine if users and professionals
use different vocabularies when describing videos, we investigate the overlap
between all user tags and a typical domain thesaurus used by professionals
in the cataloging process. A significant part of the non-verified tags — not
entered by at least two different users — are not found in the either of the
vocabularies we consider. To understand if these tags are just gibberish or
actually have meaning we perform additional experiment using the Google³
search engine as semantic filter: we deem a tag as meaningful only if the
number of pages returned by Google is positive. The procedure is motivated
by the intuition that if a person has used a word or a phrase on the Web
then it probably has some meaning. Subsequently, to shed more light on this
potentially useful class of tags we select samples from both the tags found
and not found by Google for further inspection.

With respect to the second part, we take a combined approach. First, we
investigate what do users tend to describe more: things heard or things seen
on screen. To this end we perform a study on the overlap between the user
tags and the audio signal — subtitles for hearing impaired persons — for a
sample of episodes. To get a more comprehensive understanding of the types
of tags users usually add, we perform a qualitative study of a sample of user
tags obtained through the Waisda? video tagging game. In the course of
the study each tag is manually analyzed in the light of the video content it
describes and categorized in terms of the classification framework described
in section 2.2.3.

2.4 Materials

In this section we describe the materials and resources used in the study.

2.4.1 Waisda? data snapshot

Subject of our analysis is the data collected in the first pilot project with
Waisda?, a period starting from the launch date in May 2009 until 6th of
January 2010. During this period, the game amassed over 46,000 unique tags

³http://www.google.com
2.4. MATERIALS

ascribed to approximately 600 videos by roughly 2,000 different players\(^4\). The number of distinct tag entries exceeded 420,000. The database of the game contains information about players, games, videos, and tag entries. Each tag entry is represented by an instance of a ternary relation that relates the player that entered the tag, the video the tag was attached to, and the tag itself. Additionally, a tag entry is associated with the point in time — relative to the beginning of the video — when the tag was entered. It also includes a score computed taking into consideration agreement with other tag entries in the temporal neighborhood. Since almost all players originate from the Netherlands and all videos subjected to tagging are in Dutch, the language of the vast majority of tags — nearly 100% — is Dutch.

2.4.2 Domain and lexical vocabularies

For this study we used two vocabularies: GTAA and Cornetto. While the former is a domain vocabulary, the latter is a general lexical source that covers common lexical terms.

GTAA (Dutch acronym for Common Thesaurus Audiovisual Archives) is the thesaurus used by professional cataloguers in the Sound and Vision documentation process. It contains approximately 160,000 terms divided in six disjoint facets: subjects or keywords (\(\approx 3,800\) terms), locations (\(\approx 17,000\) terms), person names (\(\approx 97,000\) terms), organization-group-other names (\(\approx 27,000\) terms), maker names (\(\approx 18,000\) terms) and genres (113 terms). GTAA terms are interlinked with each other and documented using four properties: Broader Term, Narrower Term, Related Term and Scope note. While all GTAA terms may have related terms and scope notes, only terms from subject and genres facet are allowed to have narrower and broader terms. Complementary to the narrower/broader term hierarchy, terms from the subject facet are classified by theme in 88 subcategories which are organized into 16 top-level categories.

Cornetto is a lexical semantic database of Dutch that contains 40K entries, including the most generic and central part of the language. It is build by combining Dutch Wordnet (DWN) with Referentie Bestand Nederlands (RBN) which features FrameNet-like information for Dutch [128]. Cornetto organizes nouns, verbs, adjectives and adverbs into synonym sets called synsets. A synset is a set of words with the same part of speech that can be interchanged in a certain context. Synsets are related to each other by semantic relations — like hyperonymy, hyponomy, meronomy etc. — which may be used across part of speech. Although Cornetto contains 59 different

\(^4\)Throughout this text we use the terms player and user interchangeably.
CHAPTER 2. GAME TAGS IN AUDIO-VISUAL COLLECTIONS

<table>
<thead>
<tr>
<th>Episode</th>
<th>All tags</th>
<th>Verified</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmer seeks wife 1</td>
<td>25,965</td>
<td>5,837</td>
<td>Amusement</td>
</tr>
<tr>
<td>Farmer seeks wife 2</td>
<td>22,792</td>
<td>6,153</td>
<td>Amusement</td>
</tr>
<tr>
<td>Traceless</td>
<td>1,007</td>
<td>274</td>
<td>Amusement,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Informative</td>
</tr>
<tr>
<td>Reporter</td>
<td>403</td>
<td>73</td>
<td>Informative</td>
</tr>
<tr>
<td>The Walk</td>
<td>257</td>
<td>45</td>
<td>Religious</td>
</tr>
</tbody>
</table>

Table 2.1: Sample of waisda? episodes used in the experiments.

kinds semantic relations, hyperonymy and hyponomy are by far the most frequent ones, accounting for almost 92% of all semantic relation instances.

2.4.3 Videos

For the manual classification the number of programs in the Waisda? is too large to include all of them. In addition, subtitles are not available for all videos. Therefore, for the manual classification and comparison with the subtitles we opted to use a subset of videos. We selected five episodes: the two best-tagged videos, one averagely tagged video and two low-tagged videos. The two best-tagged videos are episodes from a popular Dutch reality show, Farmer seeks Wife5, categorized as amusement. The averagely tagged video is an episode from the Traceless6 series, classified as amusement and informative program. The two low-tagged videos are episodes from The Walk7 and Reporter8 series, categorized as religious and informative, respectively. Table 2.1 summarizes the most pertinent information about the episodes. Prior research [5] suggested that the program genre might in fact influence the types of tags users add. To account for this phenomenon, we made sure that videos and fragments of all genres are present in our sample.

2.4.4 Subtitles

For the comparison of the tags with the audio signal we make use of the subtitle files associated with the television programs. Subtitles are textual versions of the dialog in films and television programs, usually displayed at

5http://www.bzv.kro.nl/
6http://spoorloos.kro.nl/
7http://dewandeling.kro.nl/
8http://reporter.kro.nl/
2.5 EXPERIMENTS

In this section we present the results from the three experiments: matching tags to vocabularies, matching tags to subtitles and manual classification of the tags.

2.5.1 Matching tags to vocabularies

In this experiment we matched all waisda? tags to two vocabularies: the general lexicon of Dutch language Cornetto and the domain thesaurus GTAA. In mapping the tags to concepts we take the following approach. We deem a tag and GTAA term to be a positive match only if they are the same string (ignoring case). A tag and Cornetto synset are considered a positive match only if at least one of the words associated with the synset is equal (in case-insensitive manner) with the tag.

The results of the mapping of Waisda? tags against Cornetto and GTAA are presented in table 2.2. We observe that only a small part of the unique tags are found in GTAA (8%). A larger number of the tags are found in Cornetto (23%). This difference between the overlap with GTAA and Cornetto is larger for the verified tags. Almost 44% of the verified tags is found in Cornetto, whereas only 14% is found in GTAA. In other words, at least 30% of the verified tags are proper Dutch words but would not be used by a

<table>
<thead>
<tr>
<th></th>
<th>All tags</th>
<th>Verified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>46,792</td>
<td>12,963</td>
</tr>
<tr>
<td>In GTAA</td>
<td>3,850</td>
<td>(8%)</td>
</tr>
<tr>
<td></td>
<td>1,825</td>
<td>(14%)</td>
</tr>
<tr>
<td>In Cornetto</td>
<td>10,939</td>
<td>(23%)</td>
</tr>
<tr>
<td></td>
<td>5,669</td>
<td>(44%)</td>
</tr>
</tbody>
</table>

Table 2.2: Overlap of Waisda? tags with GTAA thesaurus and Dutch linguistic database, Cornetto.

the bottom of the screen\(^9\). Each dialog excerpt is accompanied with time-points — relative to the beginning of the video — when the dialog excerpt appears on and disappears from the screen. The subtitles files we use were obtained from KRO broadcasting and are specified in the SubRip text file format\(^10\).

\(^9\)Timed Text Working Group, \url{http://www.w3.org/AudioVideo/TT/}

\(^10\)\url{http://en.wikipedia.org/wiki/SubRip\#SubRip_text_file_format}
CHAPTER 2. GAME TAGS IN AUDIO-VISUAL COLLECTIONS

<table>
<thead>
<tr>
<th>Facet</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>1199</td>
</tr>
<tr>
<td>Location</td>
<td>613</td>
</tr>
<tr>
<td>Genre</td>
<td>52</td>
</tr>
<tr>
<td>Person</td>
<td>118</td>
</tr>
<tr>
<td>Maker</td>
<td>4</td>
</tr>
<tr>
<td>Name</td>
<td>673</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Types</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>7222</td>
</tr>
<tr>
<td>Verb</td>
<td>2090</td>
</tr>
<tr>
<td>Adjective</td>
<td>1693</td>
</tr>
<tr>
<td>Adverb</td>
<td>171</td>
</tr>
</tbody>
</table>

Table 2.3: Waisda? tags distribution over GTAA facets and Cornetto synset types.

Professional cataloguer\(^{11}\). In addition, we observe that the verified tags are more often valid Dutch words than the non-verified ones.

Using the overlap with the vocabularies we can also provide a first classification of the tags. Using the different facets in GTAA we can distinguish different types of tags, such as subject terms, locations, persons and organization names. In WordNet we can distinguish the tags matching with different types of words, such as noun and verb. Table 2.3 shows the distribution of user tags over the GTAA facets and Cornetto synsets. We observe that most tags are matched with subject terms from GTAA, but also a large number of tags could be matched to locations and names. The overlap with Cornetto shows that most tags are matched to nouns. Surprisingly, there is also a substantial number of tags matched with adjectives. In fact, one of the most frequently occurring tags is the adjective *nice*.

From the total number of tags in total 41% are either verified or found in one of the vocabularies. Figure 2.1 provides a detailed view of these tags, showing the overlap between the different sets. We observe that 35% of the verified tags are not found in either GTAA or Cornetto. Further investigation revealed that some of these tags do correspond to terms from the vocabularies, but were not found by the matching algorithm. We also observe that 32% (the sum of 22.5%, 3.5% and 6%) of the tags are found in

\(^{11}\)GTAA contains all terms used to annotate videos in Sound and Vision
2.5. EXPERIMENTS

Figure 2.1: Overlap among verified Waisda? tags, tags in Cornetto, and tags in GTAA.

The majority of the tags, approximately 59%, are neither found in Cornetto and GTAA nor they are verified. Further analyses revealed that almost half of these tags are comprised of more than one word. While this could to some extent explain why they were not found in Cornetto and GTAA (these vocabularies predominately have single words) and they were not verified (likelihood of reaching a tag agreement among players decreases as the length of the tags increases) we still do not know if they are, in fact, meaningful. To get an answer to this question, we perform additional analysis using Google as semantic filter. For each tag we carried out a phrase search (tag was enclosed in quotes, "") and observed the number of hits (pages) that were returned. A tag is deemed as meaningful only if the number of hits returned is positive.

For approximately 84% of the tags, that were not found verified or not found in a vocabulary, Google returned positive number of hits. We sampled 200 tags from the group with no hits (zero-sample) and 200 tags with the group with positive number of hits (pos-sample) for further analysis. We discovered that the tags in the zero-sample could be divided in three groups:
garbled text with no meaning whatsoever, seriously mistyped words (bordering to garbled text), and entire sentences or excerpts from sentences mostly grammatically incorrect. The pos-sample, on the other hand, contained morphological variations of proper words, proper words combined with characters that are not letters, slang, names, idioms and phrases, and other common collocations.

In conclusion, the difference between the overlap with GTAA and Cornetto indicates that the user tags complement the vocabulary used by professional cataloguers. The tags that are found in GTAA are predominantly subject terms, but also include locations and names. We also found evidence that user agreement filters out sloppy tags, as the verified tags are more often valid Dutch words than the non-verified ones. However, a large part of the non-verified tags could still be potentially useful, as some of them can be found in GTAA or Cornetto. Moreover, the majority of non-verified tags were ‘deemed’ meaningful by Google.

It should be noted that besides case-insensitive string comparison we also considered more approximative methods such as comparison based on word stems and lemmas. The former are result of applying the linguistic procedure known as stemming and the latter are result of the linguistic procedure called lemmatisation. We also carried out string comparison based on the following string distance measures: Hamming distance, Levenshtein distance, and q-gram distance. We opted for low-level of dissimilarity since we did not want to get too many incorrect matches. Each of these string comparison techniques yielded slightly different numbers compared to the numbers reported above and given by Figure 2.1. However, the general conclusions about the low overlap between the user tags on one side and GTAA and Cornetto on the other remained the same.

2.5.2 Tags in subtitles

In this experiment we investigate the fraction of Waisda? tags that refers to the audio portion of the video content. To this end, we compare the tags associated with the five videos described in section 2.4.3 against the respective video subtitles for hearing impaired persons (see section 2.4.4).

Prior to running the analysis, all dialog text from the subtitles was broken up into words and punctuation through a process known as tokenization. Afterwards, to account for morphological variants, all words were reduced to their canonical forms through a linguistic procedure called stemming. Subsequently, the stem of each tag associated with the aforementioned videos was compared against all words in the subtitles in the appropriate video that appear at most 10 seconds before the tag was entered. The time interval of
2.5. EXPERIMENTS

<table>
<thead>
<tr>
<th>Episode</th>
<th>Tags in subtitles</th>
<th>Verified tags in subtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmer seeks wife 1</td>
<td>8,645 (33%)</td>
<td>2,546 (43%)</td>
</tr>
<tr>
<td>Farmer seeks wife 2</td>
<td>8,004 (35%)</td>
<td>2,967 (48%)</td>
</tr>
<tr>
<td>Traceless</td>
<td>182 (18%)</td>
<td>64 (23%)</td>
</tr>
<tr>
<td>Reporter</td>
<td>91 (23%)</td>
<td>16 (22%)</td>
</tr>
<tr>
<td>The Walk</td>
<td>59 (22%)</td>
<td>18 (40%)</td>
</tr>
</tbody>
</table>

Table 2.4: Overlap between the Waisda? tags and the video subtitles.

10 seconds was chosen as a reasonable amount of time needed by an average player to type in a tag. An identical time interval was used by the designers of Waisda? as the time frame for matching tags added by different players.

The results of the analysis are summarized in table 2.4. On average 26% of all tags also occur in the subtitles. This number is slightly higher when it comes to verified tags, on average 35% of all verified tags are found in the subtitles. We explain the large overlap by the fact that the audio stream of the video provides an easy way for the players to score points. This practice may, however, impair the richness of the user tags. In addition, when the subtitles of a video are available for retrieval, the user tags provide less added value.

2.5.3 Tag classification

In this experiment we performed a manual qualitative analysis on the tags of the five videos described in section 2.4.3. We only consider the verified tags of the videos. Due to the prohibitively large number of tags, for the episodes of Farmer seeks Wife we only consider the tags of two fragments. We excluded 182 tags from the sample since they were words with no descriptive power, such as particles and prepositions. In total the tag sample consisted of 1354 tags.

The tags were collectively analyzed by the authors. Each tag was considered in the light of the video fragment it describes. First, the tags were classified according to the different levels of abstraction: non-visual, perceptual and conceptual. We found no tags at non-visual level, and there were only 11 tags at perceptual level, all referring to colors. The rest of the tags (1,343) were all conceptual. The vast majority of these conceptual tags, precisely 1,313, were describing objects, whereas only 30 were about scenes. We continue our investigation by focusing on the conceptual object tags.
CHAPTER 2. GAME TAGS IN AUDIO-VISUAL COLLECTIONS

(a) Keyframe extracted from Farmer seeks Wife episode’s shot in which Yvon (the young lady) gives Amsterdam sausage as present to Anna (the elderly lady).

(b) Example of how tags (descriptions) of the keyframe above can be classified in terms of the Panofsky-Shatford model.

Figure 2.2: Classification of user tags.

In classifying the conceptual object tags we followed the guidelines compiled by Hollink et. al [52] — figure 2.2 shows an example of a classification of tags for one video fragment. We consider a tag to be specific if it possesses the property of uniqueness, for example the name of a person (Anna). A tag is abstract if its level of subjectivity allows for differences in opinion, for example, “kind lady” or “idyllic countryside”. We deem a tag to be general when only everyday worldly knowledge is required to apply it in the context of the video, for example, “woman” or “present”. To determine the facet a tag belongs to, we used the following guidelines. A tag is in the who facet if it refers to the subject (person, object, etc) of the video fragment. A tag belongs to the where facet if it refers to a location, and to the when facet if it refers to time. A tag is associated with the what facet if it refers to an object or event in the video.

Table 2.5.3 shows the distribution of the object-level tags across the cat-
2.6. CONCLUSIONS AND DISCUSSION

Table 2.5: Distribution of the object-level tags across the categories of the Panofsky-Shatford model.

<table>
<thead>
<tr>
<th></th>
<th>Abstract</th>
<th>General</th>
<th>Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>10</td>
<td>166</td>
<td>177</td>
</tr>
<tr>
<td>What</td>
<td>73</td>
<td>563</td>
<td>12</td>
</tr>
<tr>
<td>Where</td>
<td>0</td>
<td>68</td>
<td>8</td>
</tr>
<tr>
<td>When</td>
<td>4</td>
<td>31</td>
<td>6</td>
</tr>
</tbody>
</table>

Looking at the total number of tags at the different abstraction levels, we observe that the majority of the tags are general (74%), while only 7% are at the abstract level and 9% at the specific level. On the other hand, looking at the total number of tags in the facets, we observe that the majority of the tags belong to the What facet (57%). Furthermore, a considerable number of tags are in the Who facet and only a small number of tags belong to the Where and When facets. Looking at the relations between the abstraction levels and the facets we observe that almost all tags in the What facet are general, sometimes abstract, but rarely specific. The descriptions in the Who facet are, however, at both the general and the specific level, but rarely abstract. Most of the tags in the Where facet are generic, and little are specific place or country names. Finally, we encountered 195 tags that we could not classify in any of the facets. Most of the time, these tags were modifiers — typically adjectives and adverbs — that describe how an action was performed, for example nice, better, etc.

Our results show similarities with classification of image annotations by Hollink et al. [52]. They also found that a large majority of the descriptions are at the conceptual level. She, however, found a larger number of scenes (30%) at the conceptual level. A possible explanation for this difference could be the fast pace of the game, which makes the player focus on the directly perceivable objects instead of the overall scene. The evaluation of the tags by a professional cataloguer also suggested that the users focus on what can be directly seen or heard. Hollink et. al also found the majority of the descriptions to be at the general level (74%).
2.6 Conclusions and Discussion

In this section we summarize the main observations from our experiments and discuss to what extent the tags collected with Waisda? fulfill the aims of the Netherlands Institute for Sound and Vision. In addition, we discuss how the results of our study can improve future versions of the game.

From the comparison of the tags with the terms from the GTAA thesaurus of the institute and the linguistic database of Dutch, Cornetto, we made several observations. We can confirm that the aim of the institute to collect metadata in a user vocabulary can be achieved with the Waisda? video labeling game. Comparable to the results that were found in the Steve.museum tagging project, we found small overlap with the terms in the vocabulary used by professional cataloguers. In addition, almost half of the verified tags are valid Dutch words, as they were found in Cornetto.

The number of verified tags found in Cornetto is much higher than the number of tags that are not verified. This provides evidence for the assumption of video labeling games that user agreement on tags can be used to filter out ill-formed ones. We also observed that a large part of the tags that are not verified could still be potentially useful. A large part of the non-verified tags could also be found in the vocabularies. In addition, we deemed most tags meaningful as they returned results from Google.

Together with The Netherlands Institute for Sound and Vision we prepared a second pilot project with Waisda?. The results of this study show several limitations of the current metadata, that we aimed to address in the second pilot. One limitation is the low number of specific type of tags in the who and where facets. We investigated how users can be motivated to provide such tags. We showed that by matching the tags to controlled vocabularies we can derive the type of the tags. This can be used within the game to detect what type of tags are entered, and for example provide more points when the user enters a location name. For this purpose the recall of the current algorithm to match tags and terms should be improved. More details about this can be found in Chapter 5.

The manual classification of the tags provides details about the type of tags that were collected in Waisda? and how they relate to the video content. Users predominately describe what appears in the video using generic tags. Although the tags also provide some coverage of the subject, the who, and the location, the when in the video fragments. While the persons occurring as the subject are described both in generic and specific tags, there are very few tags describing specific locations.

Another characteristic of the current Waisda? tags is that many are also found in the subtitles. In case the subtitles are also available for retrieval
2.6. CONCLUSIONS AND DISCUSSION

This can be considered a limitation of the tags, as it reduces the added value. Computing the overlap between the tags and the subtitles during the game can be used to detect such tags, and for example be used to motivate users to provide different tags.

An assumption of the ESP labeling games is that only the verified tags are associated to the content as metadata. Our study shows that this approach is restrictive and would exclude many potentially useful tags. A solution could be to include tags that can be matched with a term from a controlled vocabulary. Another solution could be to compare the syntactically different tags based on their semantic similarity.

In the following chapters we will look into the usefulness of the tags in search tasks. From the current results we learned that tags describe what users directly see or hear in the video. They rarely provide a topical description of a fragment. We expect that the user tags are, therefore, suited to find objects within a specific video, but are as of yet less useful to find fragments which are about particular topic. In Chapter 4 we explore methods to detect topical descriptions of video scenes and exploit them for topical fragment retrieval.
Chapter 3

Game Tags for Visual Video Search

In this chapter we start investigating the usefulness of the game tags for search. Throughout this thesis we will address two prominent search scenarios that arise in practice. The first scenario is retrieving video fragments that feature visual appearances of objects. In fact, this is very much like the instance search task from TRECVID\(^1\) with the difference that query is formulated in text and not by visual example as in TRECVID. The second scenario we consider is topical search i.e. retrieval of video fragments that are about a given topic. In this chapter we study how effective are the game tags in the addressing the first scenario. The methodology that we use is quantitative system evaluation [127]. In order to evaluate effectiveness of information retrieval, this methodology requires collection of “documents” (in our case video fragments tagged by users via Waisda?), set of queries which we derived from real-life query logs, and relevance judgments indicating which “documents” in the collection should be returned for each query. The query set that is created in this chapter is reused in the subsequent Chapter 4, where we address the second scenario.

This chapter was published as “An Evaluation of Labelling-Game Data for Video Retrieval” in the proceedings of the European Conference of Information Retrieval 2013 [37]. It was co-authored with Michiel Hildebrand, Jacco van Ossenbruggen, Lora Aroyo, and Guns Schreiber.

\(^1\)http://trecvid.nist.gov/
CHAPTER 3. GAME TAGS FOR VISUAL VIDEO SEARCH

3.1 Introduction

Games with a purpose are a way to make humans solve tasks in an entertaining setting. Video tagging games—a type or GWAPs—could become an attractive alternative (or enhancement) to professional annotators in terms of both price and scale. While game tags are virtually for free and plentiful, professional annotations are costly and scarce. The Institute for Sound and Vision (S&V)\(^2\) launched *Waisda?*, a multi-player video labelling game where players describe streaming video by entering tags and score points based on temporal tag agreement. The underlying assumption is that tags are faithful descriptions of the videos when entered independently by at least two players within a given time-frame. From here on we shall refer to such mutually agreed upon tags as *verified* tags.

The archive expects that tags collected with *Waisda?* will improve video search. In this study, we put this hypothesis to the test. Knowing that other types of video metadata will also be present, our first research question is:

RQ1 Can game tags, on their own or in combination with other types of metadata, improve video search?

To test the assumption that agreement is a good filter, our second research question is:

RQ2 Does limiting only to verified game tags give better video search performance than considering all game tags?

When GWAPs are used to tag large video collections generally care must be taken to ensure ‘fair’ distribution of game-time across the collection items. In this sense it is instructive for collection administrators and scheduling algorithms designers to know if search performance deteriorates or stagnates after certain point, or if more tags always give better search performance. Therefore, our last research question deals with search performance change over time:

RQ3 How does the game tag search performance change when tags are added?

The rest of the chapter is structured as follows. After discussing related work, Sect. 3.3 presents our approach. Section 3.4 describes the datasets and resources that are used in our study. Section 3.5 introduces the experimental setup. Finally, Sect. 3.6 and 3.7 present the results and conclusions of this study, respectively.


\(^3\)At the time of writing, *Waisda?* is an ongoing project for three years and the game has seen its second release, [http://woordentikkertje.manbijthond.nl/](http://woordentikkertje.manbijthond.nl/).
3.2 Related Work

User annotations for video. Video annotation is a tedious and time-consuming activity. Not surprisingly, various initiatives exist that aim at collecting video annotations through crowdsourcing. In particular, LabelMe video is an online video annotation system that allows users to identify objects and annotate visual features such as motion and shapes [137]. However, this frame-by-frame conceptually low-level annotation remains a tedious task. The willingness of people to participate without compensation is limited at best. To alleviate this, [126, 105] employ the crowdsourcing MTurk platform to recruit annotators which are paid for the task. An alternative way to motivate people is to gamify the annotation experience through GWAPs. GWAPs are computer games, in which people, as a side effect of playing, perform tasks computers are unable to perform. The main example of a GWAP is Luis von Ahn’s ESP image labeling game [121]. Evaluation shows that these labels can be used to retrieve images with high precision and are almost all considered as good descriptions in a manual assessment. The idea to annotate through GWAP has been applied to video in, for example, the Yahoo! video tag game [119], VideoTag\(^4\), PopVideo\(^5\) and Waisda\(^?\). With some slight differences, in each of these games players describe streaming video by assigning free-text tags. Thus, we deem Waisda\(^?\) as a typical representative of video GWAPs.

Relevance judgements and search. Designing ground truth in the form of document relevance w.r.t. a given topic has been playing a central role ever since Cranfield experiments gained prominence[127]. The leading actor in IR benchmarking is TREC\(^6\) which employs substantial manpower in creating the ground truth. For organizations lacking the manpower, crowdsourcing is an alternative; [64, 3] showed that this task can be reliably fulfilled by crowd workers. Alternatively, Eickhoff et al. gamified the task resulting in increased reliability and reduced cost [28]. In this study we also rely on the crowd; the relevance assessment is outsourced to targeted fan groups. Search based on user-generated metadata, in particular folksonomies, has been studied before. Morrison compared the search performance of folksonomies from social bookmarking web sites against search engines and subject directories [84], showing that search engines had the highest precision and recall rates. Folksonomies, however, performed surprisingly well. Geisler and

\(^4\)http://www.videotag.co.uk/
\(^5\)http://www.gwap.com/gwap/gamesPreview/popvideo/
\(^6\)http://trec.nist.gov/
Burns state that YouTube tags provide added value for search, because 66% of them do not appear in other metadata [35]. Hildebrand et al. proposed and investigated a semi-automatic process of assigning explicit meaning to game tags for video by linking them to concepts from the Linked Open Data cloud [48]. To the best of our knowledge, no work has been done to evaluate the performance of GWAP data for video search. Our study aims to fill this void.

### 3.3 Approach

In order to assess the added value of game tags for video search we use a quantitative system evaluation methodology [127], for which we need a document collection (in our setting video fragments tagged by users in Waisda?), a set of representative queries, and relevance judgements. We created this evaluation dataset as follows: (i) select a collection of video fragments tagged by players in Waisda?, (ii) select a set of user queries from real-life query logs, and (iii) create relevance judgements. All these steps are described in more detail in Sect. 3.4.2. The dataset is then used in two experiments. In the first, we compare performance of search based on different types of metadata. On high-level the work-flow for the experiment is as follows. We create a number of systems (search engines), run them against the evaluation dataset and compute retrieval performance metrics to assess the search effectiveness of the systems. The structure of all systems is the same: each system uses the same state-of-the-art probabilistic ranking function BM25 [93], but indexes pairwise different combinations of metadata types. Thus, the only variations among the systems within an experiment is the input that they exploit for search. In our experiments we built the systems using the Xapian open source search engine library 

7 [http://xapian.org/](http://xapian.org/)
8 The default parameters for the BM25 function are outlined at [http://xapian.org/docs/bm25.html](http://xapian.org/docs/bm25.html)
ence between the search performances will stem solely from the metadata. In other words, by comparing the performance metrics of the two systems we can deduce whether using all tags for search is better than using only the matched ones. To get a better idea of what goes wrong when searching using game tags, we also carry out a follow up analysis of a sample of the results deemed as false positives. In the second experiment, we study the search performance of game tags over time. To this end, we again create a number of systems, but this time each system indexes a weekly snapshots of all tags. Comparing the performances of the systems that index consecutive weekly snapshots reveals how the tag search performance varies over time.

3.4 Datasets and Resources

In this section we describe the datasets and resources that are used in the study.

3.4.1 The MBH Video and Metadata Collection

At the time of writing, Waisda? is used to tag fragments from the popular Dutch TV program ‘Man Bigt Hond’ (MBH) produced by the Dutch broadcaster NCRV. MBH is a humoristic TV show that focuses on trivial, everyday news and ordinary, unknown people. Every episode consists of 7 or 8 unrelated, self-contained fragments where each fragment topically comes under a recurring heading\(^9\). Players in Waisda? tag these fragments. The entire collection to which we have access has 11,109 fragments from episodes aired from 2000 until 2011.

In addition to the video fragments, we have access to four types of descriptive metadata that are used as input for search:

\textbf{Waisda? tags.} We consider the collection of all game tags acquired with Waisda? during the first five months, starting from October, 2011. In this period 436,456 different tag entries were assigned to 2,192 video fragments by roughly 24,000 players. The number of unique game tags exceeds 47,000. Each tag entry is associated with the point in time — relative to the beginning of the fragment — when the tag was entered. Additionally, each tag entry is marked as ‘verified’ or not based on the tag agreement in its temporal neighbourhood. As the game is advertised only in Dutch media and the

\(^9\)The complete list of the recurring headings (categories) can be found at \url{http://www.manbijthond.nl/rubrieken}. Note that at the time of writing the webpage is in Dutch only.
material being tagged is exclusively in Dutch, the language of almost all tags is Dutch. The average number of tags per video is 199. Approximately 55% of all game tags ($\approx 243,000$) are ‘verified’ and the number of unique verified tags is 12,861. The average number of verified tags per video is 111.

**NCRV tags.** NCRV, the broadcaster, maintains an in-house collection of tags to facilitate web access to MBH fragments via search and browsing. In contrast with $Waisda?$ tags, NCRV tags are not time-based, meaning they are not linked to a particular time-point in the video, and generally cover only the prevalent topics. The average number of NCRV tags per video is 11. Thus they are usually much scarcer than the game tags.

**NCRV catalogue data.** Along with the curated NCRV tags, each MBH fragment has a short textual description, usually one paragraph, and a title. We consider the collection of all titles and textual descriptions (i.e. catalogue data) as another metadata type that will be used in the study.

**Captions.** Closed captions are textual versions of the dialogue in films and television programs for the hearing impaired, usually displayed at the bottom of the screen. Each dialogue excerpt is accompanied with time-points — relative to the beginning of the video — when the dialogue excerpt appears on and disappears from the screen. We use captions obtained from S&V that cover most of the MBH episodes aired in 2010 and 2011 which amounts to a total of 897 fragments.

### 3.4.2 Evaluation Dataset

In this section we describe the creation of the three separate components of our evaluation dataset: set of video fragments, set of queries, and relevance judgements.

**Video fragment subset.** The set of fragments for our experiment is selected from the MBH fragments tagged in $Waisda?$. Not all metadata types described above are available for every single fragment. To do a fair comparison of the search performance of various metadata types, we use only a subset. The filtering criterion is as follows: we include only the fragments that have at least one $Waisda?$ tag and NCRV tag ascribed to them, and for which captions files are available. This results in a collection of 197 fragments. The accumulative duration of our test collection is almost 11 hours of
3.4. DATASETS AND RESOURCES

Figure 3.1: Number of tags contributed by each of the players in the selected sample of fragments. The horizontal axis represents the players ordered decreasingly on the number of tags they entered. The vertical axis displays the logarithm to base 2 of the number of tags entered by the associated player.

video material, with an average fragment length of approximately 3.3 minutes and a median of 3.6 minutes. The duration of the shortest and the longest fragment in our collection is 0.5 and 8.6 minutes, respectively. The total number of game tags, verified game tags, and NCRV tags ascribed to the videos of these collection is 107,531, 80,805, and 2,066, respectively. Thus, the average number of game tags, verified game tags, and NCRV tags per fragment is 545, 410, and 10 respectively. Figure 3.1 displays the number of tags contributed by each of the players in the selected sample of fragments. The number of players that entered at least one tag for one of the fragments in the sample is 473. As can be seen from the figure there is a small core of power taggers which contributed significant portion of the tags. In fact, each of the top ten players has entered more than 500 tags. The median number of tags entered by a player for the sample of fragments is 28.

Query set. To measure the information retrieval performance we use real-life user queries. NCRV provided us with one month of query logs from the MBH web site. The logs contain 15,219 queries posed by the internet users to the site’s search engine asking for video fragments. Figure 3.2 shows the query frequency distribution. As seen, the query frequency follows a power law; aside from few frequent ones most of the queries appear infrequently. In fact, only 6% of the queries appear at least 5 times (points under or on the lower horizontal dashed line in Figure 3.2).
Out of the complete set of 15,219 user queries we select, in two steps, a subset of 50 queries to include in the study. First, we partition the query set into three classes: a high, mid and low frequency class. The borders of the classes are chosen so that the area under the curve in Figure 3.2 for each class is one third of the area. Queries appearing more then 22 times form the high-frequency class, between 5 and 22 form the middle-frequency class, and queries appearing less than 5 times form the low-frequency class.

Second, for each class we perform filtering. Namely, a query is skipped whenever it meets one of the following criteria: (i) it equals the title of one of the MBH recurring headings or it contains one of the words ‘man’, ‘bijt’, and ‘hond’ from the series title. (ii) it is not found in at least two of the metadata types described in Sect. 3.4.1. We consider the first criterion in order to exclude the non-informative queries\(^\text{10}\) and we consider the second criterion to avoid creating bias towards any metadata type. After the filtering, we are left with 12, 78, and 49 queries from the high-frequency, middle-frequency, and low-frequency class, respectively. The top 12, top 19, and top 19 queries from the high-frequency, middle-frequency, and low-frequency class, respectively, comprise the final query set. All queries in the query set are single-term queries. This is a consequence of our second selection criterion — the requirement of a query being present in at least two metadata types which reduces the probability of multi-term queries being selected. The second se-

\(^{10}\)Queries that are expected to have a low inverse document frequency (IDF)
collection criterion was introduced mainly to be fair towards NCRV tags and catalog data. Our analysis of the query logs revealed that generally queries were present either in the Waisda? tags or the captions, sometimes both. In contrast, the queries from the query logs were less frequently present in the NCRV metadata. Thus, if we omit this criterion — and potentially include multi-term queries — our expectation is that it will not drastically change the outcome for the Waisda? tags and the captions. However, the NCRV metadata will likely perform worse.

**Relevance judgements.** In order to collect relevance judgements for the query set and the fragment collection we performed an on-line user experiment. To this end, we deployed a web application which was used by the participants to carry out the evaluation. For each participant the workflow proceeds as follows. Whenever a participant accesses the web application she is presented with a welcome page which contains a description of the task she is required to perform. Before starting with the evaluation, the participants need to fill out a questionnaire that aims at assessing their familiarity with Waisda?, the MBH TV series and the MBH website. Then the participants proceed to the evaluation page (see Figure 3.3) which plays a randomly assigned fragment and lists the complete query set. During the evaluation process, the participants watch the fragment and indicate which of the concepts denoted by the queries are shown in it. We asked users to judge a fragment to be relevant for a query if it depicts the concept denoted by the query. Each participant is asked to evaluate at least five fragments.

**Participants**

The participants in the experiment were recruited mainly from the Waisda? online community and MBH series fanbase by distributing a call for participation through the major social networking services Facebook\(^{11}\) and Twitter\(^{12}\). The posted messages and tweets contained a link to our web application. 107 participants started the experiment, 83 of them evaluated at least one fragment and 25 participants evaluated more than 5 fragments. Judging from the questionnaire data, the level of familiarity of participants with MBH series almost uniformly ranges from ‘never seen it’ to ‘watch it regularly’. Surprisingly, the participants who never visited the MBH website or visited it only few times are the vast majority. Also for familiarity with Waisda?; the participants who never played or played only few times are the overwhelming majority.

\(^{11}\)http://www.facebook.com/
\(^{12}\)https://twitter.com/
Figure 3.3: Screenshot of the evaluation page. At the top, a video player is placed which displays the fragment. The list of queries is rendered at the bottom.

**Participant’s (Dis)agreement**

From the entire collection of 197 video fragments, 134 of them are evaluated by 2 distinct participants. The rest of the fragments, 63 in total, are evaluated by 3 distinct participants. When consolidating the relevance judgements from different participants we use majority voting; the side — either ‘relevant’ or ‘not relevant’ — that gets more votes wins. In case of a tie, we take the side of ‘relevant’ i.e. we deem the fragment to be relevant for the query. We justify this decision with the following reasoning. The notion of relevance in our particular case is defined in terms of depiction of the concept denoted by the query in the fragment. Our queries are not abstract concepts and there is very little room for different interpretations among the participants. Thus, we believe if one participant rated a query ‘not relevant’ and another ‘relevant’ for a given fragment it is most probable that the first participant simply missed it. The consolidated evaluation set is publicly available online at https://goo.gl/AQcpGz. The overlap among the participants in terms of evaluated videos is too small to reliably measure the inter-rater agreement with measures such as Krippendorff’s alpha. However, we found that the probability of a rater rating ‘relevant’ is 9.5% and the probability of disagreement between raters is 10.1%.
3.5 Experiments

To answer the research questions formulated in Sect. 3.1 we use a quantitative system evaluation. Namely, we implement a number of search engines and run them against the evaluation dataset described in Sect. 3.4.2. In all experiments we evaluate the performance of the various search engines using the mean average precision (MAP) measure. The number of results returned by the systems is low enough (not more than 30) for the users to be willing to inspect them all. Thus, we deem that it is important that all results are good not just the top ones. This intuition is captured by MAP. To assess if the difference in performance is statistically significant we use the student’s paired t-test at 0.01 level of significance as suggested by [103].

3.5.1 Experiment 1

In this experiment we address the first and the second research question. To this end, we retrieve fragments for the set of queries using 12 search engines. Each of the search engines utilizes the same state-of-the-art probabilistic ranking function BM25 and the only variation among them is the data they index. Consequently, differences in retrieval performance are attributed solely to the data. We implement search engines that index:

1. \(SE_{user}\) all Waisda? tags
2. \(SE_{vuser}\) only verified Waisda? tags
3. \(SE_{ncrv}\) all NCRV tags
4. \(SE_{catalog}\) NCRV catalogue data
5. \(SE_{caps}\) all captions
6. \(SE_{caps+user}\) all captions and all Waisda? tags
7. \(SE_{caps+catalog}\) all captions and all catalogue data
8. \(SE_{ncrv+caps}\) all captions and all NCRV tags
9. \(SE_{ncrv+user}\) all NCRV tags and all Waisda? tags
10. \(SE_{baseline}\) all NCRV tags and catalog data
11. \(SE_{all-except}\) all metadata except Waisda? tags
12. \(SE_{all}\) all metadata types including Waisda? tags

\(SE_{baseline}\) is an approximation of the search functionality offered on the web site dedicated to MBH series. We use it as a baseline for comparing the search performance of the other search engines. By comparing the performance of \(SE_{user}\) and \(SE_{vuser}\) we are able to see if using all tags as opposed to only verified tags is detrimental or beneficial for fragment search (\(RQ2\)). Furthermore, comparing the performance of \(SE_{user}\) and systems 3 through 12 will reveal how well game tags are doing — on their own and in combination — compared to other types of metadata (\(RQ1\)).
3.5.2 Experiment 2

In this experiment we address the third research question. We retrieve fragments for the set of queries using two collections of search engines. The first collection consists of search engines that index snapshots of all game tags taken periodically once a week. Identically, the second collection consists of search engines that index snapshots of the verified tags taken at the same time points as the snapshots from the first collection. As with experiment 1, all search engines use the same probabilistic ranking function BM25 and the only variation among them is the data that they index. Examining the performance of search engines within a collection reveals how tag search performance changes over time. By examining the performance of search engines across collections we learn how all tags perform compared to verified tags.

3.6 Results

In this section we present the results of our experiments.

3.6.1 Experiment 1

The results for this experiment are summarized in Table 3.6.1. As seen, considering only verified tags yields worse search performance than considering all tags. Intuitively, verified tags should yield higher precision but lower recall than all tags. Indeed, the average search precision of verified tags (0.59) across the queries is higher than the average search precision of all tags (0.49). However, search based on all tags yields more relevant results — the average search recall of all tags (0.42) is higher than the averages search recall of the verified tags (0.28). In fact, for 35 queries the non-verified tags yielded relevant results — on average 4—not found by verified tags. It seems the tag verification criterion is too conservative in a sense that it filters out tags that are in fact useful for search.

Search based on game tags ($SE_{user}$) significantly outperforms search based on other metadata types alone. Indeed, search based on game tags is approximately 69% more successful than search based on the in-house NCRV tags ($SE_{ncrv}$). We believe this is attributed to the fact that NCRV tags are relatively scarce and cover mainly prevalent topics. In this sense, game tags are complementary to the NCRV tags and the combination of both is mutually beneficial. Indeed, the search engine that indexes both game tags and NCRV tags

---

13 A snapshot contains all game tags up to a given point in time.
14 More detailed figures can be found the Appendix A at the end of the thesis.
3.6. RESULTS

<table>
<thead>
<tr>
<th>$SE_{user}$</th>
<th>$SE_{vuser}$</th>
<th>$SE_{ncrv}$</th>
<th>$SE_{catalog}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.219↑†</td>
<td>0.143↑†</td>
<td>0.138↓‡</td>
<td>0.077↑†</td>
</tr>
<tr>
<td>$SE_{caps}$</td>
<td>$SE_{caps+user}$</td>
<td>$SE_{caps+catalog}$</td>
<td>$SE_{ncrv+caps}$</td>
</tr>
<tr>
<td>0.157↑‡</td>
<td>0.247↑†</td>
<td>0.183↑‡</td>
<td>0.201↑†</td>
</tr>
<tr>
<td>$SE_{ncrv+user}$</td>
<td>$SE_{baseline_{ncrv+catalog}}$</td>
<td>$SE_{all−user}$</td>
<td>$SE_{all}$</td>
</tr>
<tr>
<td>0.263↑†</td>
<td>0.150↑‡</td>
<td>0.208↑‡</td>
<td>0.276↑†</td>
</tr>
</tbody>
</table>

Table 3.1: Results for experiment 1: MAP scores for the search engines — MAP score for given search engine is given immediately below. †, ‡, and ≈ indicate if a score is significantly better, worse, or statistically indistinguishable from the MAP scores of $SE_{user}$ and $SE_{vuser}$, in that order.

tags, $S_{ncrv+user}$, yields a performance increase of 20% and 90% over search engines $S_{user}$ and $S_{ncrv}$, respectively.

Furthermore, search based on solely on game tags yields better performance from our baseline search engine, $SE_{baseline_{ncrv+catalog}}$. Indeed, the MAP scores of $S_{user}$ and $SE_{baseline_{ncrv+catalog}}$ indicate a performance increase of 46%.

Comparison of the MAP scores of $SE_{user}$ and $SE_{caps}$ indicates that game tags outperform captions by approximately 39%. This can be explained by the fact that captions only cover the audio portion of the video content, whereas game tags cover both audio and visual. In fact, previous work [39] suggested that players tend to describe more things that appear visually in a video.

Combination of captions and game tags proves to be beneficial: $SE_{caps+user}$ outperforms $SE_{caps}$ and $S_{user}$ by 64% and 13%, respectively.

Lastly, the search engine that indexes all available types of metadata, $SE_{all}$, performs best. This is to a large extend due to the contribution of game tags. Indeed, $SE_{all}$ outperforms the search engine that indexes all metadata types except for game tags, $SE_{all−user}$, by 33%. Obviously, the said difference can only be attributed to the effect of the game tags. Interestingly, search based on game tags alone outperforms by 5% $SE_{all−user}$, which is the best performing search engine that does not index game tags.

Analysis of False Positives

One of the conclusions drawn above is that the precision of $SE_{user}$ — or in other words when considering all game tags — is relatively low. In order to get a better insight into the reasons that lead to poor precision we manually
analysed a sample of the false positives\footnote{A search result is a false positive if it is returned by $SE_{user}$ but was deemed ‘not relevant’ by our evaluators.} among the results obtained from $SE_{user}$. Note that given the definition of search precision, it is the false positives that are responsible for the drop in the precision score.

In the course of the analysis, we randomly picked fifteen queries from the query set and investigated the false positives retrieved for them — 192 in total. For each query-fragment pair we watched 10 seconds prior and 10 seconds after the point when the tag that matched the query was entered. The occurrence of the query instance in the fragment is then classified as either ‘visual’, ‘nonvisual’ based on the same criterion used by the evaluators, or unrelated. The unrelated category consists of tags that are not describing the content of the video and can be accounted as errors by the users. The ‘visual’ class is further subdivided into the following disjoint subclasses which designate how/where the instance denoted by the query is depicted along the spatial and temporal dimensions:

- **Intro.** The instance is in the opening segment of the fragment which is usually unrelated with the rest of the fragment.
- **Short-lived in background.** The instance appears for a brief moment, e.g. less than a second, in the background and not in opening segment. An example of this would be a car passing by in the background.
- **Short-lived in foreground.** The instance appears for a brief moment, e.g. less than a second, in the foreground and not in opening segment. An example of this would be a person passing by in front of the camera.
- **Long-lived.** The instance appears either in the background or in the foreground for a longer period of time.

The ‘nonvisual’ class is subdivided into the following disjoint subclasses:

- **Captions in intro.** The instance appears only in captions in the opening segment.
- **Captions.** The instance appears only in captions and not in the opening segment.
- **Screen text.** Screen text referring to the instance.

The results of the classification are summarized in Table 3.2. As seen, significant fraction of the false positives were caused from tags associated with the
opening segments: 17% of them refer to visual instances and 12% are mentioned in the captions in the opening fragment. Moreover, 15% of the false positives are a result of the tags found only in the captions not in the opening segments. In general, it is context-dependent whether these cases should be considered relevant or not. In this analysis, we deemed visual appearances in opening segments as relevant and references in captions as not relevant due to our interpretation of (visual) relevance. However, in other scenarios the interpretation of relevance may differ. Nevertheless, it is straightforward to detect these cases automatically and include/exclude them as input for search depending on the context. The short-lived background/foreground and other cases were most likely missed by the evaluators. Had they been correctly identified the outcome of the comparison given above would have been even more favourable for Waisda? tags. Furthermore, 20% of the false positives were visual appearances of long-lived instances i.e. objects appearing on screen for a longer period of time. However, it is interesting to mention that these cases usually involved ambiguity and the context and the hints provided by the current or the previous scenes from the video were required to make the correct interpretation. For example, for the query boer (the Dutch word for farmer) there were instances where a person while being a farmer, a fact established in the previous scenes, in one particular scene was not in the typical farmer’s attire and could not be identified as such using only the information provided on the screen. It is easy to imagine how a tagger having seen the video from the beginning and having this extended context could interpret the person as a farmer. Needless to say, what is the right answer in these cases depends on the particular application context. Finally, 21% of the false positives were caused by incorrect tags not at all related with the content of the fragments.

3.6.2 Experiment 2

In this section we present the results from our second experiment which addresses the third research question. Figure 3.4(a) shows the MAP scores of the search engines indexing the weekly snapshots of all tags and only the verified tags. Figure 3.4(b), on the other hand, shows how the number of all tags and verified tags increased over time. Looking at Figure 3.4(a) we conclude that most of the time the search performance for both the verified tags and all tags is monotonically increasing with the number of tags. In other words, the more tags we amass, the better our effectiveness in searching fragments becomes. Furthermore, looking at the pairwise search performance differences between the search engines that index the weekly snapshots of all tags and verified tags (vertical dashed lines between plots in Figure 3.4(a)),
CHAPTER 3. GAME TAGS FOR VISUAL VIDEO SEARCH

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intro</strong></td>
<td>33 (17%)</td>
</tr>
<tr>
<td><strong>Visual</strong></td>
<td></td>
</tr>
<tr>
<td>Short-lived in background</td>
<td>13 (7%)</td>
</tr>
<tr>
<td>Short-lived in foreground</td>
<td>15 (8%)</td>
</tr>
<tr>
<td>Long-lived</td>
<td>38 (20%)</td>
</tr>
<tr>
<td><strong>Captions in intro</strong></td>
<td>23 (12%)</td>
</tr>
<tr>
<td><strong>Nonvisual</strong></td>
<td></td>
</tr>
<tr>
<td>Captions</td>
<td>28 (15%)</td>
</tr>
<tr>
<td>Screen text</td>
<td>1 (0%)</td>
</tr>
<tr>
<td><strong>Unrelated</strong></td>
<td>41 (21%)</td>
</tr>
</tbody>
</table>

Table 3.2: Experiment 1 false positives classification.

we conclude that using all tags for search opposed to only verified ones yields consistently better results. In fact, search performance improvements are statistically significant for every single pair.

The performance of search based on all game tags surpasses our baseline, $SE_{nrcv+catalog}$, around the 11th week after 42,271 tags have been collected (Figure 3.4(b)). Beyond that point the said difference in performance steadily increases as more tags are collected. With $SE_{all-user}$, which is the best performing search engine that does not index game tags, this happens a bit later. In particular, after the 18th week and 91,508 collected tags, $SE_{user}$ starts to outperform $SE_{all-user}$. Thus, there is a point somewhere between the 18th and 19th week when the collected game tags outperform all search engines that do not index tags.

It is also interesting to note that the precision and recall of search based on all tags are monotonically non-decreasing with the number of tags for each query in our set. The results for search precision and recall can be found in the appendices B and C at the end of the thesis, respectively.

3.7 Conclusions

In this chapter we have studied the added value of game tags for video search. For this reason we have created a publicly available evaluation dataset that consists of real-life user queries, a video fragment collection, and relevance judgements.

Search based solely on game tags outperforms search based on other types of metadata such as in-house (NCRV) tags or captions. Thus if any of the
other metadata types are unavailable or costly to acquire, relying only on sufficient game tags for search could yield equal or even better results. In our dataset, combining game tags with other metadata types is beneficial for search. In fact, the search engine that exploits all available metadata performs best, to large part due to the contribution of the game tags—the observed performance improvement is 33%.

Exploiting only verified game tags for search gives poorer performance than search based on all game tags. While search based on verified tags yields higher precision, it also has lower recall compared to all game tags. In fact, for most of the queries non-verified tags provided relevant results that were not found by the verified tags. This proves that considering only verified tags is too conservative filtering criterion resulting in discarding non-verified game tags that are valid video descriptors and thus useful for search.

Search performance steadily increases as more game tags are collected. This is true for both verified and all tags. Moreover, search based on all tags consistently outperforms search based only on verified tags. When the average number of tags is slightly more than 2 tags per second, the search using all tags outperforms all search engines that are not indexing game tags. Such an estimate could be used as an indicator whether a video has been tagged enough.

In Chapter 3 we established that game tags usually describe events and objects within scenes so it is not surprising that they excel at visual search. We also established that there is a difference in the focus and the scope of the game tags compared to the professional annotations; game tags being more fine-grained and professional annotations being more coarse-grained referring to the entire video and describing the prevalent topics. The in-house catalogue tags fit the latter profile: they are scarce and cover the topics of
the video. This being said, quite predictably they fall short compared to the
game tags when it comes to visual search. The closed-captions, on the other
hand, are similar to the game tags in terms of focus and scope. Yet, game
tags outperform them. This can be explained by the fact that closed-captions
only cover the audio portion of the video and there are some aspects that
have only visual representation. However, in order for the game tags to be
able to compete with closed-captions there should be a sufficient number of
them assigned to the videos. What is a sufficient number of tags for a video
is very hard to tell but in principle the more tags the better. And with time
as more tags are added their search performance is improving and tables are
turning in their favour. This was demonstrated in Sect. 3.6.2. In the next
chapter we investigate the added value of the tags collected with Waisda?
for topical search.
Chapter 4

Game Tags for Topical Video Search

In this chapter we continue our investigation of the usefulness of the user tags for search. Particularly, our aim is to evaluate the performance of user tags for retrieval of videos that are about a given topic i.e. topical search. To this end, we will reuse the same methodology as in Chapter 3, namely quantitative system evaluation [127]. The results presented bellow demonstrate that raw, unprocessed game tags are not well suited for retrieving video fragments based on topic. While the search recall is satisfactory, the search precision leaves much to be desired. This is mainly caused by the presence of game tags which are not valid topical annotations. Thus, the second aim in this chapter is to characterize the quality of the user tags as topical annotation and to detect and filter out the non-topical ones. We explore several features of the tags which could serve as an indication of their quality as topical descriptors. Our results show that after filtering, game tags can emulate the retrieval performance of a baseline system that utilizes manually crafted metadata for search. An important consequence of this finding is that tagging games can provide a cost-effective alternative in situations when manual annotation by professionals is too costly.

This chapter is an extension of a journal article which was published as “Topical Video Search: Analysing Video Concept Annotation through Crowdsourcing Games” in the International Journal of Human Computation 4(1), 47-66, 2017 [38]. It is co-authored with Michiel Hildebrand, Jacco van Ossenbruggen, Lora Aroyo, and Guus Schreiber.
CHAPTER 4. GAME TAGS FOR TOPICAL VIDEO SEARCH

4.1 Introduction

In the past decade audio-visual (AV) content collections have been undergoing a transformation from archives of analogue materials to very large stores of digital data accessible online\(^1\). As the AV collection items become accessible to the wider internet audience the lack of adequate annotations is highlighted: users cannot find what they are looking for because annotations are either not present or too expert-centric — created from the perspective of the catalogers [87]. Video tagging games are an attempt to alleviate this problem. By engaging the internet community to tag their videos, the AV collection owners can benefit in at least two important ways. First, there is a clear cost-benefit, collecting annotations in this way is usually cheaper than using professional documentalists. Second, the game tags can help bridge the terminological gap in search if the searchers and annotators originate from the same community and use similar terminology when searching and tagging.

One of the most successful ongoing online video tagging games is *Waisda?*. The game was launched in 2009 by the Netherlands Institute for Sound and Vision (S&V), one of the largest AV archives in Europe\(^2\). The first pilot of the game run until January 2010 and produced over 420,000 user tags. Considering the acquired experience and insights, S&V launched a second version of the game (see Chapter 5 for more details) which features several improvements, albeit the basic idea of the temporal tag agreement is preserved.

The second version of *Waisda?*\(^3\) is a mature, production-grade, customizable GWAP with diverse usage. It is included\(^4\) as a showcase in the Europeana\(^5\) internet portal (see Section 5.4.1 for more details). *Waisda?*

---

\(^1\)An excellent example is PrestoPRIME, \url{http://www.prestoprime.org/}, which was an European Union funded project about digitisation and preservation of the AV heritage. The consortium included the major national AV archives in Europe.

\(^2\)S&V manages and preserves over 70% of the Dutch AV heritage. The collection contains more than 750,000 hours of television, radio, music and film from the beginning in 1898 until today.

\(^3\)At the time of writing the second version of the game has been discontinued. S&V is deploying (still under development) a third installment of the game available at \url{http://waisda.beeldengeluid.nl/}. S&V intends to integrate *Waisda?* more tightly in their internal workflows and to use it to collect tags for the items in their online collection \url{http://in.beeldengeluid.nl/}.

\(^4\)\url{http://labs.europeana.eu/apps/waisda-floss}

\(^5\)Europeana is an internet portal that provides multi-lingual access to millions of books, paintings, films, and archival records that are part of the European cultural heritage. More than 2,000 institutions across Europe have contributed to the project.
has also been deployed and incorporated in Spotvogel\(^6\) (see Section 5.4.2 for more details). In the former the internet users tag archival footage from the European cultural heritage and in the latter they identify occurrences of wildlife in footage. These are two relatively different domains yet \textit{Waisda}? was fitted to them seamlessly. This is due to the simple design of the game and the unconstrained way the tags are entered — the players get a small set of instructions and then are free to enter whatever they please. This is a double-edged sword. On the one hand, the entry barrier for the players is low and the target audience is wide as no specific skills are needed. On the other hand, the open-ended way the tags are entered may be overwhelming in the fast-paced setting of the game and players may succumb to entering low quality tags which lack descriptive power just to reach consensus with other players and win points. In this study we perform a qualitative analysis to investigate this issue.

The cost-effectiveness of \textit{Waisda}? versus manual annotation by professionals comes with a caveat: successful deployment of a GWAP is no small feat. Attracting new players and keeping them engaged over time is vital for success and requires continuous publicising efforts. The empirical evidence collected from both \textit{Waisda}? pilots supports this claim. On both occasions the bulk of the tags was accumulated during periods when there was an active campaign for promoting the game. Targeting the fanbase of the TV series via various channels (e.g. TV series website, social media, etc.) has proved to be a successful strategy for attracting new players. Player’s engagement is sustained with in-game mechanisms such as leaderboards to honour the best players and motivate the others, and with time-limited contests offering awards for the top performers. Planning and executing activities like these is certainly costly, however in the case where the growth rate of AV collections is ever increasing, these costs will be outweighed by the costs of manual annotation. This is one of the lessons learned from the \textit{Waisda}? project.

The second version of the game has amassed more than 710,000 user tags. The number of unique tags is 71,448 and each of the top five most-tagged videos has more than 3,600 tags ascribed to it which amounts to an average tag density of more than 13 tags per second. This is a substantially higher number than the number of tags assigned by professional catalogers, typically 10-20 tags per video. However, are all these user tags of any use? Does this

\(^6\)Spotvogel (http://spotvogel.vroegevogels.vara.nl/) which translates to mocking bird from Dutch, was a \textit{Waisda}? deployment by the Dutch broadcaster VARA (http://www.vara.nl/) and the S&V institute with the aim of collecting user tags for the footage from the Dutch television program Vroege Vogels (http://vroegevogels.vara.nl/). The game run for 6 months in 2013 and was nominated for Dutch Game Awards 2013 http://www.dutchgameawards.nl/2013/spotvogel/, note that the link is in Dutch.
overwhelming quantity implies quality? In Chapter 3 we demonstrated that the Waisda? tags can indeed be successfully exploited for retrieving video fragments that feature visual appearances of objects of interest (persons, objects, etc.) [37]. Our results showed that Waisda? tags excel at this kind of visual instance search, outperforming the closed captions and annotations created by professionals.

In this study we wish to understand whether the fast pace of Waisda? forces players to have tunnel vision about the content of the video and tag predominately non-topical aspects. Previous analysis performed by a senior cataloger from S&V ruled the game tags to be limited in scope, referring mainly to things seen or heard on the screen [6]. However, we hypothesize that players do describe topics and topical tags are present albeit buried under the myriad of non-topical tags. The first research question, therefore is

RQ1 Do players enter tags which describe the topics of the videos?

In the context of topical video search, the presence of non-topical tags, which do not refer to the topics covered in the video, can lead to false positives. For example, if a video \( v \) has a tag \( t \) ascribed to it and \( t \) does not refer to the topics covered in \( v \). If a searcher is interested in videos that are about \( t \) and uses \( t \) as a search term the video \( v \) will be incorrectly retrieved. Thus, it is important to know whether a tag is non-topical and ignore it in the retrieval process to eliminate its negative influence on the search results. Our second research question, therefore is

RQ2 Can the access to videos based on topic be improved by detecting and filtering out the non-topical game tags?

The rest of the chapter is structured as follows. In Section 4.2 we discuss related work. Section 4.3 outlines our approach. In Section 4.4.1 and 4.4.2 we describe the collection of video fragments and the evaluation dataset, respectively. Section 4.5 presents the findings with respect to the effectiveness of the user tags for topical search. This sets the baseline for the remainder of the study. In Section 4.6 we outline a number of tag filters and evaluate their effectiveness for eliminating non-topical tags. Section 4.7 presents the conclusions from the study.

4.2 Related Work

User annotations for search. Search based on user-generated metadata, in particular folksonomies, has been studied before. Morrison compared
4.2. RELATED WORK

web search performance of folksonomies from social bookmarking web sites against search engines and subject directories [84], showing that search engines had the highest precision and recall rates. Folksonomies, however, performed surprisingly well. In fact, user tags show promise to alleviate the vocabulary mismatch problem for search: bridging the gap between user queries and metadata used for retrieval [34]. Indeed, Geisler and Burns state that YouTube tags provide added value for search, because 66% of them do not appear in other metadata [35]. Heymann et al. investigated a large-scale sample of forty million bookmarks from the social bookmarking site del.icio.us and found that in 20% of the cases user tags do not occur in the page text, backlink page text, or forward link page text of the pages they annotate. Studies in [11, 43, 94, 135] investigate this phenomenon across multiple domains and multimedia resource types and identify the gaps between the tag space and the querying vocabulary. The common conclusion is that user tags can improve search by bridging the vocabulary gap.

Studies reported in [10, 110, 80] take a more critical stance. They conclude that while overall user tags improve search, not all tags are suitable for retrieval. In fact, Marshall [80] even suggests that tags may be less effective descriptors for image retrieval, classification, and description than other forms for descriptive metadata such as title and narrative captions. This hints that a characterization of the quality of the tags is needed to filter out the tags that are not suited for retrieval. This is one of the aspects that our study addresses.

Another line of research is exploiting the tripartite structure (Users $\times$ Tags $\times$ Resources) of folksonomies to improve search [53, 7]. Alternatively, the semantics of tags can be grounded in some lexical sources and the grounded tags utilized for improving search. For example, Hildebrand et al. proposed and investigated a semi-automatic process of assigning explicit meaning to user tags for video by linking them to concepts from the Linked Open Data cloud [48]. Gligorov et al. evaluate the performance of user tags collected with a GWAP for retrieval of video segments that depict particular concept of interest (person, object, etc.) [37].

Quality and Refinement of Annotations. There is a substantial body of research into the refinement and quality assessment of annotations of still images. Lee at al. propose a tag refinement technique that aims at differentiating noisy tag assignments from correct tag assignments [73]. Each tag is assigned a probability of being noisy based on the visual similarity of the images and tag co-occurrence statistics. Tags with a probability above a threshold are discarded as noisy.
In [114, 77] neighbour voting schemes for determining the tag relevance are explored. In this approach, a tag is considered more relevant to the image it is ascribed to, also known as the seed image, if the tag is also used to annotate the neighbouring images. The neighbourhood relation is defined in terms of the visual similarity among images. Lee at al. expand the approach by not only considering the visually similar images, but the dissimilar images as well, thus providing negative examples [74]. Kennedy at al. exploits visual similarity among images in a sense that tags ascribed to images by the creators of the image are used as seed annotations and also attached to visually similar images [65]. Zhao at al. propose a data-driven method to automatically determine the relatedness between a tag and the image’s visual content taking into consideration the tag co-occurrence and the visual similarity among images [139].

Probabilistic methods that exploit random walk based techniques have also been explored [129, 79, 76]. These methods produce a ranking of the tags according to their relevance with respect to the image with which they are associated. The tag relevance estimations are computed as the stationary or the convergence probabilities of a random walk processes. Notable example of this approach is the PageRank algorithm [26, 67, 63]. Another group of methods exploit background knowledge (such as the lexical database Wordnet and a massive corpus indexed by Google) to perform the refinement of the image annotations [60, 130]. The semantic relations encoded in Wordnet and the semantic similarity quantified by the Google-based measures like the Normalized Google Distance [24] provide contextual evidence for the relationship among the annotations. This evidence is then used as an input for machine learning algorithms which give the final word for the quality of the annotations.

**Games With a Purpose (GWAPs)** GWAPs are a human-based computation technique in which humans solve tasks, too difficult for computers, in a game setup which provides entertainment for the players. The predecessor of all GWAPs is the ESP game designed by von Ahn and Dabbish [121], which harnesses human abilities to label images. The general idea of the ESP game is that two players who can see the same image try to come up with matching tags. The players are paired up randomly without any means of communication. Therefore, when players agree on a tag it is a strong indication that the tag is a valid descriptor of the image. In [123] von Ahn and Dabbish generalize the design principles of the ESP game into a conceptual framework for designing GWAPs. Given a problem which is hard or impossible to solve by computer the framework provides high-level guidelines and
principles for transforming the problem into GWAP that solves it. Following
the ESP game, many GWAPs were designed that tackle problems from var-
ious domains. *Peekaboom* is a GWAP in which players identify and locate
objects in images [125]. The output of the game is precise location informa-
tion and other useful information which can be used to train computer vision
algorithms. Several GWAPs have been designed for collecting or validation
of common sense knowledge. *Verbosity* is a GWAP [124] which addresses the
problem of creating a database of ‘common-sense facts’, statements about
the world known to most people. Similarly, *Common Consensus* is a GWAP
that aims at collecting a database of human goals [78]. *Top10* and *Pirate
& Ghost* [20], on the other hand, are GWAPs that deal with the problem of
verification of the common sense assertions of the form \( \text{concept} \rightarrow \text{relation} \rightarrow \text{concept} \). Another area where GWAPs are applied is collecting visual data for
real world locations. *EyeSpy* is a GWAP in which players contribute photos
or textual data about geographic locations [8]. The collected data is sub-
jected to in-game validation by other players and once validated can be used
for navigation tasks. *PhotoCity* is another GWAP where players contribute
photos of real world locations in urban areas [115, 116]. The ultimate goal is
to use the collected photos to build detailed 3D models of real world places
varying over time. GWAPs have also been applied in the music and art do-
 mains. *TagATune* [70, 71] and *Listen Game* [117] are GWAPs where the aim
is to collect annotations for music pieces. *Artigo*\(^7\) is a game platform that
offers six GWAPs for annotating artworks [132]. Three of the games Artigo
game, Artigo Taboo, and TagATag are variations of the ESP game. TagATag
includes a more challenging aspect where players are tagging pairs consisting
of an image and a tag. The resulting tags describe relationships between the
members of the pair (the image and the tag). The other three game of the
platform, Karido, Artigo-Quiz, and Combino are designed to complement
the data collected with the first three ESP-like games. In *Karido* [108] the
process of annotation is carried out as a guessing game where one player is
trying to guess *goal image* out of set of images. Second player tries to help the
first player by providing description of the image. If the first player correctly
guesses the *goal image* then the description is valid annotation with high
probability. The aim of the Combino game is to collect semantically more
complex multi-word tags [109]. Semantic web is another area for application
of GWAPs. *OntoGame* is a series of GWAPs that aims to cover the complete
Semantic Web life-cycle: building and maintaining ontologies, alignment of

\(^7\)Artigo, [http://artigo.org](http://artigo.org), is a game platform for artwork annotation as well as artwork semantic search engine in German, English, and French founded in 2008. According to [16], thanks to a good media coverage over several years, the German version of ARTigo has a sufficient number of regular players.
ontologies, and semantic annotation of data [102]. Building and maintain-
ing ontologies is carried out with the OntoPronto GWAP [112], alignment of
ontologies is achieved with the SpotTheLink GWAP [113], and semantic an-
notation of resources is done with the OntoTube and the OntoBay GWAPs.
GiveALink Slider and Great Minds Think Alike are two more GWAPs that
address the problem of collecting reliable semantic annotations [131]. An-
other area where GWAPs are applied is linguistics. Jinx is a GWAP that
tackles the problem of word sense disambiguation [97]. Borsa Parole and
Poker parole are GWAPs that aim at collecting linguistic data [15]. In Borsa
Parole word phrases and their characteristics are collected from the user
community. Poker parole aims at collecting meta-data about the data col-
lected in Borsa Parole i.e. players form conjectures about the word phrases
and their characteristics. Other notable GWAPs are Odd Leaf Out which
addresses the problem of misclassified leaf images [45] and Polarity which
deals with collecting attributes and attribute values for resources [69]. Pearl
and Steyvers in [89] outline a GWAP-based methodology for identifying emo-
tions, intentions, and attitudes in text. Kneissl in [66] studies how GWAPs
and more generally crowdsourcing can be used in the field of e-learning.

Important area of research is on the limitations of ESP-like games. Weber
at al. demonstrated that the ESP game in its original design encourages
players to enter ‘obvious’ or predictable tags [92]. Weber at al. built a
probabilistic language model which using only the already assigned tags for
the image and without any knowledge about the image was able to predict
with high probability 8 what the new tags added by players will be. Jain and
Parkes provided a game-theoretic explanation for this observation [57]. Their
game-theoretic model of the ESP game indicated that from the players’ point
of view it is more beneficial if they focus on more obvious (low-effort) tags.

4.3 Approach

To answer the research questions stated above we use a quantitative sys-
tem evaluation methodology [127] which requires an evaluation dataset that
consists of three components: a document collection (in our setting video
fragments tagged by players in Waisda?), a set of representative queries,
and relevance judgements. To evaluate the performance of a search system,
the query set is run against the system and the retrieved results are graded

---

8 Even without any understanding of the actual image, the probability of agreement
with randomly assigned human partner on a label was 69% for all images, and 81% for
images which have at least one tag assigned to them.
4.3. APPROACH

w.r.t the relevance judgements\(^9\). The search performance or the search system is then quantified by search performance metrics such as Mean Average Precision (MAP), Recall, Precision, etc. Central to this evaluation methodology is the creation of the relevance judgements which plays the role of gold standard. In the making of the gold standard we exploit in-house annotations created by the broadcaster which describe the main topics in the video. More precisely, we deem a video to be relevant for a query if the query concurs with at least one the topics (annotations) of the video (see Section 4.4.2).

Our study is divided in two parts. In the first part we address the first research question: do players enter topical tags and in effect how suited are the game tags for topical search. To this end, we create four search systems, each exploiting different metadata. We run each system against the evaluation dataset and carry out comparative analysis of their retrieval performance. The first and the second system use all game tags and verified only game tags, respectively. The third system is our baseline and exploits the in-house catalog metadata — it is an approximation of the present search functionality. The forth system combines all game tags and the in-house catalog metadata. By comparing the first and the second system against the baseline system, we deduce how suited the game tags for topical search are on their own, and see if they could replace the need for the current expensive cataloguing practice. By comparing the forth system with the baseline, we infer whether the game tags provide added value on top of the existing catalog metadata. To get a better qualitative insight of what happens under the hood, we carry out an analysis of example false/true positives. We classify the game tags according to the Panofsky-Shatford scheme \(^{52}\) to establish which parts of the video content are described by the tags. The aim is to derive insights about users’ tagging practices that are associated with ascribing topical tags.

In the second part we address the second research question: we investigate several ways to detect and filter out non-topical tags with the goal of improving topical search performance. In particular, we take tag features such as TF-IDF score and player’s reputation and derive a binary yes/no decision as to whether to retain the game tag or drop it. Each filtering method yields a subset of all game tags. To judge how well each of the filtering methods work we build a search system for each of them which uses the corresponding subset of all game tags as input for search. Moreover, to see how well the filtering methods work together with the existing catalog metadata we

\(^9\)Relevance judgments give binary (yes/no) answer as to search result is relevant to the query. Thus, the retrieval process was fully automatized.

\(^{10}\)Tags entered by at least two players within a time interval of 10 seconds.
build a search system for each of them which uses the corresponding subset of all game tags and the catalog metadata as input for search. The search performances of these systems are compared to the baseline system from the first part of the study (see Sect. 4.6 for more details).

When comparing the performance metrics of any two systems, to assess whether the difference is statistically significant we use the student’s paired t-test at 0.01 level of significance as suggested by [103].

4.4 Experimental Data

In this section we describe the data and the evaluation dataset used in this study. More precisely, Section 4.4.1 outlines the metadata that we exploit for search and Section 4.4.2 describes in more detail the evaluation dataset.

4.4.1 The MBH Video and Metadata Collection

In the second pilot, Waisda? was used to tag fragments from the popular Dutch TV program ‘Man Bijt Hond’ (MBH, English: ‘Man Bites Dog’) produced by the Dutch broadcaster NCRV\(^{11}\). MBH is a humoristic TV show that focuses on trivial, everyday news and ordinary, unknown people. Every episode consists of 7-8 unrelated, self-contained fragments where each fragment topically comes under a recurring heading. Players in Waisda? tag these fragments. The entire video collection to which we have access has 11,109 fragments from episodes aired in the last 11 years.

In addition to the video fragments, we have access to four types of descriptive metadata that we use as input for search in this study:

Waisda? game tags. We consider the collection of all user tags acquired with Waisda? during the first five months, starting from October, 2011. In this period 436,456 different tag entries were assigned to 2,192 video fragments by roughly 24,000 players. The number of unique user tags exceeds 47,000. Each tag entry is associated with the point in time — relative to the beginning of the fragment — when the tag was entered. Additionally, each tag entry is marked as ‘verified’ or not based on whether the tag was entered by at least two players within a time interval of 10 seconds. As the game is advertised only in Dutch media and the material being tagged is exclusively in Dutch, the language of almost all tags is Dutch. The average number of tags per video is 199. Approximately 55% of all user tags (≈ 243,000)

\(^{11}\)http://www.ncrv.nl/
4.4. EXPERIMENTAL DATA

are ‘verified’ and the number of unique verified tags is 12,861. The average number of verified tags per video is 111.

NCRV tags. NCRV, the broadcaster, maintains an in-house collection of tags to facilitate web access to MBH fragments via search and browsing. In contrast with Waisda? tags, NCRV tags are not time-based, meaning they are not linked to a particular time-point in the video, and generally cover only the prevalent topics. The average number of NCRV tags per video is 11. Thus they are usually much scarcer than the game tags.

NCRV catalog data. Along with the curated NCRV tags, each MBH fragment has a short textual description, usually one paragraph, and a title. We consider the collection of all titles and textual descriptions (i.e. catalog data) as another metadata type that will be used in the study.

Captions. Closed captions are textual versions of the dialogue in films and television programs for the hearing impaired, usually displayed at the bottom of the screen. Each dialogue excerpt is accompanied with time-points — relative to the beginning of the video — when the dialogue excerpt appears on and disappears from the screen. We use captions obtained from S&V that cover most of the MBH episodes aired in 2010 and 2011 which amounts to a total of 897 fragments.

4.4.2 Evaluation Dataset

In this section we describe the three components of our evaluation dataset: set of queries, set of video fragments, and relevance judgements. The evaluation set is publicly available online at https://goo.gl/zXZeYq.

Query set

We will reuse the same set of queries from a previous evaluation of the Waisda? tags for visual search [37]. Here by visual search we mean keyword search where the goal is to retrieve videos that visually depict the object/artifact of interest. The query set consists of fifty queries that were sampled from user query logs of the TV series’ web site. The sampling procedure involved grouping the queries into three classes based on their frequency in the logs: high, mid, and low frequency class. Each of the three classes were then filtered to remove bias and to ensure fair comparison of the retrieval performance of the metadata types. The top-ranked 12, 19, and 19, queries, w.r.t. frequency, from the high, mid, and low frequency class, respectively,
comprise the final query set. The exact details of the sampling procedure can be found in [37]. An example of a high-frequency query is *Mandy* which is the main character in one of the rubrics in the MBH show. On the other hand, an example of a low-frequency query is *Friesland* which is a region in the Netherlands and the filming location of one of the rubrics in the MBH show.

An objection may be raised for reusing the same queries from the visual search study, where the goal is to retrieve videos that *depict the artefact* specified by the query, for topical search, where the goal is to retrieve the videos that are *about the topic* specified by the query. However, most queries can refer both to an object/artefact appearing in the video or to one of the topics the video is about. For example, consider the query *horse*. A video may depict a horse as part of the scenery and this would qualify it as a relevant result for visual search where we might be looking for stock footage containing horses. Another video may be about equines and therefore be a relevant result for topical search on the topic *horses*. In fact, the query selection procedure used in the visual search study [37] was oblivious about the searcher’s intent behind the queries, be it visual, topical or something else. Only after the queries were selected a visual search interpretation was assigned to them and the gold standard was created accordingly. In the same manner, in this study we give the queries a topical interpretation which is embodied in the gold standard (see Sect. 4.4.2 below). Figure 4.1 shows the number of topically relevant videos per query as judged by the gold standard. As seen, for every query there is positive number of relevant videos which means all queries can be interpreted topically.

**Video fragment set**

The set of fragments for this particular experiment is selected from the MBH fragments tagged in *Waisda?*, described in more detail in Section 4.4.1. For a fair comparison of the search performance, we select a subset from the entire collection of fragments. The selection criterion is as follows: only fragments that have at least one verified *Waisda?* tag ascribed to them are considered. The resulting collection contains 2,562 fragments with accumulative duration of almost 123 hours of video material. The average fragment length is approximately 2.9 minutes and the median is 3.2 minutes. The duration of the shortest and the longest fragment in our collection is 0.1 and 25 minutes, respectively. The total number of user tags, verified user tags, and NCRV tags ascribed to the videos of this collection is 591,468, 355,522, and 28,248, respectively. Thus, the average number of user tags, verified user tags, and NCRV tags per fragment is 231, 139, and 11 respectively.
Ground Truth

As we said earlier, NCRV tags are in-house tags that describe the topics of the fragment with which they are associated. We consider the NCRV tags as the ground truth about what topics are covered by the fragments. With this in mind, given a query \( q \) and a fragment \( f \), we deem \( f \) to be topically relevant for \( q \) if there is an NCRV tag that is equal with \( q \), a synonym of \( q \), or a hypernym of \( q \). \( Y \) is a hypernym of \( X \) if every \( X \) is a \( Y \), for example canine is a hypernym of dog. Hyperonymy is a transitive relation [82]. More formally, the topical relevance relation is defined as follows

\[
\text{Topical Relevance} = \{(q, f) \mid \exists t \in NCRV(f) \land \left( lower(q) = lower(t) \lor \right. \\
\left. \text{synonym}(lower(q), lower(t)) \lor \left. \text{hypernym}(lower(t), lower(q)) \right)\}\}
\]

(4.1)

where \( NCRV(f) \) is the set of all NCRV tags associated with the fragment \( f \), \( lower(\cdot) \) is the lower case string function, \( \text{synonym}(w_1, w_2) \) is a binary predicate which is true iff \( w_1 \) and \( w_2 \) are synonyms, and \( \text{hypernym}(w_1, w_2) \) is a binary predicate which is true iff \( w_1 \) is a hypernym of \( w_2 \). Figure 4.1 shows the number of relevant fragments in our collection for the queries in the dataset along with additional descriptive statistics. As seen, the median and the average number of relevant hits for a query is approximately 27 and 34, respectively. The lowest number of relevant hits for a query is 10 whereas the highest is 113.

The NCRV tags were primarily created to be used for searching and
browsing on the TV-show website by the online community. The tags are displayed prominently on the website’s UI which enhances the social proof effect [33, 40], thus many of the tags are incorporated in the terminology of online community. In fact, our chosen query set belongs to the intersection between the NCRV tags and the online community’s search terminology as witnessed by Fig. 4.1. As implied by the definition (4.1) and seen in Fig. 4.1, for every query in the query set there is at least one video annotated with the same NCRV tag. This is sufficient for our narrower aim to determine the relevance for this particular query set and not for any query in general. In other words, the mismatch between the NCRV tags and online community’s terminology has no impact on the derived relevance judgements for our query set.

Note that while the NCRV tags are currently used by the broadcaster on the TV-show website, there is no guarantee that they, when used as described above, form a complete and accurate ground truth for topical search. We choose this set-up because the alternative scenario, creating a dedicated topical search ground truth for a limited subset ourselves, would have resulted in a much smaller data set.

We considered variation of definition 4.1 where we defined the topical relevance relation only by considering case-insensitive string comparison (omitting synonyms and hyponyms). Even with the modified definition the general conclusions from Sections 4.5.1 and 4.6 below remained the same; while the retrieval metrics of the systems (given by Tables 4.1 and 4.5) varied slightly the ordering did not change and the differences remained statistically significant.

4.5 Evaluation of Game Tags for Topical Video Search

In this section we will evaluate the effectiveness of the Waisda? tags for topical video search. The results obtained here will serve as a starting point for comparison in the subsequent sections where we will try to improve the retrieval effectiveness of the Waisda? tags by filtering those that are irrelevant as topical descriptors.

4.5.1 Game tags vs. catalog data

To address the first research question stated above we created four search engines. Each of them utilizes the same state-of-the-art probabilistic ranking function BM25 and the only variation among them is the metadata they
4.5. **EVALUATION OF GAME TAGS FOR TOPICAL VIDEO SEARCH**

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SE_{user}$</td>
<td>0.131†</td>
<td>0.16†</td>
<td>0.589†</td>
</tr>
<tr>
<td>$SE_{vuser}$</td>
<td>0.081‡</td>
<td>0.193‡</td>
<td>0.286‡</td>
</tr>
<tr>
<td>$SE_{catalog}$</td>
<td>0.168††</td>
<td>0.438††</td>
<td>0.291‡†</td>
</tr>
<tr>
<td>$SE_{user+catalog}$</td>
<td>0.151‡‡</td>
<td>0.17‡‡</td>
<td>0.654‡‡</td>
</tr>
</tbody>
</table>

Table 4.1: MAP/Precision/Recall scores for the search engines. †, ‡, and ≈ indicate if a score is significantly better, worse, or statistically indistinguishable from the MAP scores of $SE_{user}$ and $SE_{vuser}$, in that order.

We index and use as input for search. Consequently, differences in retrieval performance are attributed solely to the data. We evaluated search engines that index:

1. $SE_{user}$ all Waisda? tags
2. $SE_{vuser}$ only verified Waisda? tags
3. $SE_{catalog}$ NCRV catalog data
4. $SE_{user+catalog}$ catalog data and all Waisda? tags

We did not consider using captions for search because we only have them available for a small subset of MBH fragments. If we would have used them in the study, we would have had to settle for much smaller evaluation collection of fragments. Naturally, we did not use the NCRV tags either, since we used them as ground truth for topical relevance.

As said, the combinations of metadata types that are indexed by the various systems are strategically chosen so that the resulting performance metrics from the evaluation dataset will provide answers to the first research question. We compare the performance of $SE_{user}$ against $SE_{catalog}$ to evaluate the performance of the game tags alone when compared to the catalog metadata. By comparing $SE_{user}$ against $SE_{user+catalog}$ we see if the performance of the catalog metadata can be further improved by adding the game tags. Furthermore, we check whether the performance of the game tags can be improved by just using the verified tags ($SE_{user}$ versus $SE_{vuser}$).

4.5.2 **Results**

Table 4.1 summarizes our findings; the detailed precision and recall metrics for each query can be found in appendix D. Compared to the previous study (reported in Chapter 3) where we looked at visual relevance, the results for topical relevance are disappointing. The search performance of game tags ($SE_{user}$) is 28% below that of the NCRV catalog data ($SE_{catalog}$). Combining the game tags with the catalog metadata does not help either: $SE_{catalog}$ is better than $SE_{user+catalog}$ by 11%. The good news is that $SE_{user}$ outperforms
The results are even worse for the verified user tags. While they yield only marginally better search precision, their recall is far worse (see Table 4.1). Consequently, \( SE_{\text{user}} \) outperforms \( SE_{\text{vuser}} \) by 62% which suggests that considering all tags is better for topical search than limiting the scope only to verified tags.

In conclusion, the good recall of the game tags does not outweigh the low precision. The latter is caused by \( Waisda? \) tags that match the query but are associated with fragments that are not considered topically relevant. Below, in Section 4.6, we attempt to detect and filter out these tags to improve the search performance.

**A Closer Look on Verified Tags**

The results presented in the previous section suggest that verified tags leave much to be desired when it comes to topical search. To get a better qualitative insight of what happens behind the scenes we carry out an analysis of samples of the results returned by the system that indexes the verified \( Waisda? \) tags. Our hypothesis is that the tags referring to objects that are visually depicted in the videos are the leading reason for the poor retrieval performance. To test this we analyse a subset of the false positives and a subset the true positives returned by \( SE_{\text{vuser}} \). For each query we consider all returned videos— either true or false positive — for which captions are available. The subject of the analysis is the tag that caused a given video to be returned for a given query.

First, we classify each tag in the samples into three categories based on the content component — audio, visual, or both — it refers to (note that the classification was carried out by a single rater). Tags that refer to concepts that are visually depicted but are not mentioned in the dialog are classified as only visual. Tags that are mentioned only in the dialog are classified as only in captions. The third category contains the tags that refer to concepts that are both visually depicted and present in the dialog. Second, for each of the categories we compute the average TF-IDF (Term Frequency - Inverse Document Frequency) for the tags in that category. TF-IDF is a numerical statistic which reflects how important a term is to a document in a collection or corpus [61, 95]. In our context, the TF-IDF score of a tag reflects the relevance of the tag for the associated fragment where the fragments are represented as bags of all tags ascribed to them. The average TF-IDF score of a category is an indicator of the average relevance of its members relative to the other categories. Lastly, to establish what aspects of the video content
4.5. EVALUATION OF GAME TAGS FOR TOPICAL VIDEO SEARCH

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Average TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only visual</td>
<td>361 (67%)</td>
<td>22.302</td>
</tr>
<tr>
<td>Visual and in captions</td>
<td>19 (3%)</td>
<td>55.793</td>
</tr>
<tr>
<td>Only in captions</td>
<td>161 (30%)</td>
<td>49.319</td>
</tr>
</tbody>
</table>

Table 4.2: False positives analysis.

the tags from our samples are describing we classify them according to the Panofsky-Shatford model [52]. This model divides the descriptions into three levels: general (generic things in the video), specific (specific things), and abstract (symbolic things). Each of the levels is further broken down into four facets: who, what, where, and when producing the Panofsky-Shatford 3x4 matrix. There are alternative tag classification schemes [100], however we picked Panofsky-Shatford since it provides insight about the relation of the tag and video content it describes; the role of the concept, denoted by the tag, in the video content is captured by the model, e.g. if the tag Amsterdam is classified in Where Specific this would signify that this is the location of the scene in the video.

Sampling procedure. The tags for our analysis are selected as follows. For each query we consider all returned videos for which captions are available. The sample of tags consists of all verified tags ascribed to these videos. The reason we restrict only to videos with captions is that in the course of the analysis we check for presence in captions, as described above. The results of the analysis are given in continuation.

False positives

We start by analysing the false positives returned by the system that indexes the verified tags, $SE_{user}$. The results of the analysis are summarized in Table 4.2. The total number of analysed instances is 541. As suspected, the majority of the false positives, about 67%, are caused by tags referring to concepts that are only visually depicted. Around 30% of the false positives are caused by tags that appear only in the audio component of the content. The remaining 3% are caused by tags present both in audio and visual part of the content. Interestingly, it seems that the tags which are present in the audio and refer to a concept that appears visually are less likely to yield false positives. Their presence in both the audio and visual component is a strong indication that they are denoting salient aspects of the content. This is witnessed even more by the fact that this category has the highest average TF-IDF score which measures the importance of a term for a document. In
CHAPTER 4. GAME TAGS FOR TOPICAL VIDEO SEARCH

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Average TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only visual</td>
<td>20 (27%)</td>
<td>50.195</td>
</tr>
<tr>
<td>Visual and in captions</td>
<td>33 (45%)</td>
<td>96.182</td>
</tr>
<tr>
<td>Only in captions</td>
<td>21 (28%)</td>
<td>60.185</td>
</tr>
</tbody>
</table>

Table 4.3: True positives analysis.

<table>
<thead>
<tr>
<th></th>
<th>True positives</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>General</td>
</tr>
<tr>
<td>Who</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>What</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>Where</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>When</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Classification of positives.

In this section we continue our analysis with the true positives returned by the system $SE_vuser$. The results are summarised in Table 4.3. The total number of analysed instances is 74. The figures are following the same trend as the figures for the false positives analysis documented in Table 4.2. The tags that are present in the captions and represent concepts depicted visually make the category that yields the highest number of true positives, 45%. Around 28% of the true positives are result of tags which are present only in the captions. The remaining 27% are yielded by tags that denote concepts depicted only visually. Again, the category of tags found both in the audio and visual component have the highest TF-IDF score.

We also classified the positives using the Panofsky-Shatford model [52]. More precisely, we classified the tags that led to the hit with respect to the returned video fragment. The results from the classification are summarized in Table 4.4. It is interesting to note the disproportionally large number of false positives compared to the number of true positives for the What and Where facet. In fact, the set of false positives in the What facet significantly overlaps with the only visual false positives set from Table 4.2. For most part these are objects appearing in the foreground and the background of the scenery. The case of the Where facet is the more interesting one. The false positives in the General Where facet are mostly caused by tags which refer to the dialog in situations where the actors are talking about generic places e.g. their current whereabouts like the farm. The false positives in the Specific
Where facet almost all originate from one query, namely Amsterdam. In fact, our collection features a series of videos where a TV crew from Amsterdam travels to other places in the Netherlands and interviews ordinary people. At the beginning of every interview the crew presents themselves at which point they mention that they come from Amsterdam. It is a signature motif in the series and is usually picked up by the players in Waisda?.

After analysing the true and false positives, the general conclusion is that the verified tags usually refer to the more “obvious”, more noticeable, aspects of the content which is in agreement with the conclusions from [92, 57]. For example, moving objects in the background or in the foreground, prominent stationary objects, or words from dialog are among the things the verified tags denote. This is hardly surprising, after all these are things that are easy to reach consensus on and ultimately that is the goal of the game from the perspective of the players.

4.6 Filtering Non-topical Game Tags

Previously, in Sect. 4.5 we investigated how well the tags collected with Waisda? are performing with respect to topical search. The general conclusion was that when it comes to using Waisda? tags to retrieve fragments that are about a given topic the search performance is unsatisfactory. This conclusion came with the following caveat: while the search recall is relatively high ($\approx 59\%$), the search precision is rather low ($\approx 16\%$). What this means is that a significant portion of topics covered by the fragments are in fact entered as tags by the Waisda? players hence the high recall. Moreover, there are also many tags that do not refer to the topics covered by the fragments and when these tags result in a hit the overall search precision goes down. This being said, should the non-topical tags be detected and filtered out from the collection, that would result in increased precision and unchanged recall.

Tag Filters To make our discussion more precise we introduce the notion of a tag filter. Speaking formally, a tag filter $F : \mathcal{T} \rightarrow \{true, false\}$ is a unary boolean-valued function defined over the set of all tags, $\mathcal{T}$. For example, we can express the notion of verified tag using tag filters. Indeed, a filter $F_{\text{ver}}$ which evaluates to true iff it is passed a verified tag as an argument can be defined in the following way.

$$F_{\text{ver}}(t) = \begin{cases} true & \text{if } \exists t' \in \mathcal{T} \ (t \neq t' \land v(t) = v(t') \land l(t) = l(t') \land |\tau(t) - \tau(t')| < 10) \\ false & \text{otherwise} \end{cases}$$
for every $t \in T$ where $v(\cdot)$ is a function that returns the video fragment the
tag passed as an argument is attached to, $\tau(\cdot)$ is a function that returns
the time in seconds relative to the beginning of the video when the tag was
entered, and $l(\cdot)$ is a function that returns the label of the tag$^{12}$. Given
a video fragment $v \in V$ and a tag filter $F : T \rightarrow \{true, false\}$, the set
of tags that remain after the filtering is denoted by $Filter : (V \times T \rightarrow
\{true, false\}) \rightarrow \mathcal{P}(T)$ which is defined as follows$^{13}$

$$Filter(v, F) = \{t | t \in tags(v), F(t) = true\}$$

(4.2)

where $tags(\cdot)$ is a function that returns all tags attached to the fragment
passed as an argument. For instance, $Filter(v, F_{\text{ver}})$ denotes the set of all
verified tags attached to video fragment $v$.

### 4.6.1 Tag Filters

In this section we describe the tag filters that are investigated in this study.

**Entered by at least $k$ Players**

As seen in Sect. 4.5, considering only verified tags for search results in rela-
tively low search performance. This is mainly due to the fact that verification
conditions, as formulated by $F_{\text{ver}}$ above, are too restrictive. Consequently,
we end up throwing away good tags. Our first try is to relax the conditions
by omitting the temporal agreement requirement. Topical tags refer to the
entire video, but as our previous studies suggested, the 10 seconds agreement
window does not reflect this well as it favours tags that are describing more
obvious and short-lived aspects of the content. In particular, we consider the
tags that are entered by at least $k$ players where $k \geq 2$. More formally, we
define the tag filter $F_{\text{play},k}$ as follows

$$F_{\text{play},k}(t) = \{|\{p | p \in \mathcal{P}, \exists t' \in tags(v(t)) (t \neq t' \land l(t) = l(t') \land
p(t') = p\}| \geq k\}$$

(4.3)

for every $t \in T$ where $\mathcal{P}$ is the set of players, the functions $tags(\cdot), v(\cdot)$ and
$l(\cdot)$ are the same as defined above, and the function $p(\cdot)$ returns the player
that added the tag passed as an argument. In our experiments, we will vary
the value of $k$ in the set $\{2, 3, 4\}$. Going beyond $k = 4$ drastically reduces
the number of remaining tags.

$^{12}$The label of the tag is the actual text entered by the user that contributed the tag.

$^{13}\mathcal{P}(T)$ denotes the power set of $T$
4.6. FILTERING NON-TOPICAL GAME TAGS

**TF-IDF-rank Take Top k**

We saw in Sect. 4.5.2 that the TF-IDF score of a tag is rather indicative as to whether the tag is topical or non-topical. Referring back to Sect. 4.5.2, for the analyzed sample of positives, the average TF-IDF score of the true positives was higher than the average TF-IDF score of the false positives. This suggests that the TF-IDF measure favours topical tags. Assuming the correctness of this hypothesis, we exploit this measure to filter out the potentially non-topical game tags. In particular, for each fragment in the collection we rank the tags associated with the fragment based on their TF-IDF scores in descending order: the tag with the highest TF-IDF score is at the top. Then the filtering is performed by taking only the top $k$ tags for every video. This is expressed by the tag filter $F_{\text{tfidf},k}$ defined as follows

$$F_{\text{tfidf},k}(t) = \begin{cases} \text{true} & \text{if } \text{rank}_{\text{tfidf}}(t) \leq k \\ \text{false} & \text{otherwise} \end{cases}$$

for every $t \in T$. The function $\text{rank}_{\text{tfidf}} : T \rightarrow \mathbb{N}$ returns the position (rank) of a given tag $t$ in the ranking in descending order based on TF-IDF score of the tags in the set $\text{tags}(v(t))$. As a clarification, given our previous definitions, $\text{tags}(v(t))$ denotes the set of tags associated to the fragment $v(t)$ to which the tag $t$ is associated. We vary the value of $k$ in the set $\{5k | k \in \mathbb{N}, 2 \leq k \leq 20\}$, i.e. all integers from 10 to 100 with increments of 5.

**Latent Dirichlet allocation-based filtering**

Topic models [49, 12] are a type of statistical models for discovering abstract ‘topics’ in a collection of documents. One of the most common topic models currently in use is the Latent Dirichlet Allocation (LDA). The idea behind LDA is to model documents as arising from multiple topics, where a topic is defined to be a probability distribution over a fixed vocabulary of terms — the set of all unique words in the collection. Specifically, it is assumed that there exists a fixed set of $K$ topics associated with the collection, and that each document exhibits these topics with different proportions. Furthermore, LDA assumes that words are exchangeable within each document, i.e., their order does not affect their probability under the model. In other words, each document is treated as a ‘bag of words’. We believe that the assumptions underlying the LDA model are valid in and applicable to our context as well. Videos, much like documents, have many layers of meaning and can be

---

14The TF-IDF measure is computed over a corpus of documents. In this particular case the “documents” are the bag of tags associated with the fragments.
viewed as mixture of topics. User tags collected through Waisda? can be seen as instantiations of these topics. The unstructured nature of the tags, — only weak temporal ordering of tags within video exists, based on the tag entry time — fits the ‘bag of words’ metaphor quite well. According to LDA the probability that a tag \( t \) appears in video \( v \) is given by

\[
P(t \mid v) = \sum_{i=1}^{K} P(t \mid z_i)P(z_i \mid v)
\]

where \( P(t \mid z_i) \) is the probability of the tag \( t \) for the topic \( z_i \) and \( P(z_i \mid v) \) is the probability of picking a tag from the topic \( z_i \) in \( v \) i.e. the proportion of \( z_i \) exhibited by \( v \). We use the probability \( P(\cdot \mid v) \) given by (4.5) to rank the tags ascribed to \( v \) in descending order. The filtering is carried out by taking the top \( k \) tags for \( v \). This is expressed by the tag filter \( F_{\text{lda},k} \) defined as follows

\[
F_{\text{lda},k}(t) = \begin{cases} 
\text{true} & \text{if } \text{rank}_{\text{lda}}(t) \leq k \\
\text{false} & \text{otherwise}
\end{cases}
\]

for every \( t \in \mathcal{T} \). The function \( \text{rank}_{\text{lda}} : \mathcal{T} \rightarrow \mathbb{N} \) returns the position (rank) of a given tag \( t \) in the ranking in descending order of the set \( \text{tags}(v(t)) \) based on the probability distribution \( P(\cdot \mid v) \) given by (4.5). We vary the value of \( k \) in the set \( \{5k \mid k \in \mathbb{N}, 2 \leq k \leq 20\} \), i.e. all integers from 10 to 100 with increments of 5.

**Player reputation-based filtering**

Another aspect that can be exploited for filtering is the reputation of the players. By incorporating the player’s reputation in the tag filtering process we aim to reduce the influence of the ‘bad’ players. According to [32], reputation of an entity (e.g. person or an organization) is an opinion about that entity, usually a result of some process of evaluation based on evidence about the entity. In our context, as evidence we consider the events of players ascribing tags to videos which are described by the identity of the player, the tag, the time and the id of the fragment. In fact, this is precisely what is logged by the system and nothing more. A limiting factor is the fact that most of the players (\( \approx 99\% \)) are anonymous or in other words they only have one recorded session in which they played one or more games. This means that even if the same person played two different sessions as anonymous player there is no way to reliably correlate the sessions. In effect, the amount of evidence we can collect for players is limited. We define the reputation of a player as the ratio of the number verified tags entered by the player to the number of all tags entered by the player. We consider a verified tag to be
4.6. FILTERING NON-TOPICAL GAME TAGS

a positive evidence for the player’s reliability, therefore a higher fraction of verified tags implies higher player’s reputation. Instead of computing the ratio directly, we estimate the value by taking the lower bound of Wilson score confidence interval for a Bernoulli parameter [1, 133]. The latter approach is more robust in cases where the number of observations (evidence) is small. Now that we have a way to quantify the reputation of the players we define the filtering procedure. The idea is for each tag to combine its TF-IDF score with the reputation of the player that entered it. In particular, for each fragment \( v \) we rank the tags in descending order according to the following score

\[
score(t) = \text{tfidf}(t) \times \text{rep}(p(t))
\]

for each tag \( t \) ascribed to \( v \) where \( \text{tfidf}(\cdot) \) is the function that denotes the TF-IDF score and \( \text{rep}(\cdot) \) returns the reputation for the player \( p(t) \) that entered the tag \( t \). We see that the score is proportional with the reputation i.e. higher reputation results in higher score. The filtering is carried out by taking only the top \( k \) tags for every video fragment. This is expressed by the tag filter \( F_{\text{prep},k} \) defined as follows

\[
F_{\text{prep},k}(t) = \begin{cases} 
\text{true} & \text{if } \text{rank}_{\text{score}}(t) \leq k \\
\text{false} & \text{otherwise}
\end{cases}
\]

for every \( t \in T \). The function \( \text{rank}_{\text{score}} : T \rightarrow \mathbb{N} \) returns the position (rank) of a given tag \( t \) in the ranking in descending order of the set \( \text{tags}(v(t)) \) based on the score function given by (4.7). We vary the value of \( k \) in the set \( \{5k| k \in \mathbb{N}, 2 \leq k \leq 20\} \), i.e. all integers from 10 to 100 with increments of 5.

**Network Analysis-based filtering**

Network analysis tools and mechanisms [44, 25] are increasingly used to study folksonomies and social tagging related phenomena [59, 134, 99, 81, 46, 13]. The crux of these approaches is to represent the domain knowledge as a network (graph) and apply network analysis tools to investigate the phenomenon of interest. In our particular case, we exploit network analysis to detect and filter out non-topical tags. The general idea is to build a network for each video fragment that captures the semantic connectedness among the tags associated with that fragment. Once the network is built we exploit network centrality measures to rank the tags according to their importance. The intuition is the more central a given tag is the higher its connectedness with the other tags is and therefore that tag has higher importance as content descriptor. We consider three centrality measures: pagerank, weighted
degree centrality and eigenvector centrality. However these methods did not
perform as well as TF-IDF and LDA and for space reasons we omit them
from our discussion.

4.6.2 Search Engines

As said, our goal is to investigate various ways how the non-topical tags
can be detected and filtered out with the ultimate goal of improving search
performance. To this end, we define a number of tag filters and apply them
to every fragment in the collection in the manner described above. Each tag
filter yields a filtered collection of tags which is used as an input for search
by systems (search engines). Note that there are two systems per tag filter:
one that indexes only the filtered collection of tags and one that indexes both
the filtered collection of tags and the catalog metadata. The structure of all
systems is identical: each system uses the same state-of-the-art probabilistic
ranking function BM25, but indexes different metadata for search, depending
on the tag filter and whether or not catalog data is considered. Thus, the
only variations among the systems is the input that they exploit for search\textsuperscript{15}.

4.6.3 Results

In this section we present the results of running the search engines that
index the filtered collection of tags using the tag filters described in Sect.
4.6.1 against the evaluation dataset. The evaluation metrics are summarized
in Table 4.5. For reference we copied the metrics for the systems $SE_{user}$,
$SE_{vuser}$, and $SE_{catalog}$ from Sect. 4.5.2. We use the following notational con-
vention in the table: the system that indexes the data obtained by applying
the filter $F$ is denoted by $SE_F$ and the system that indexes the data obtained
by applying the filter $F$ and the catalog data is denoted by $SE_F+catalog$. The
system that we are trying to beat is $SE_{catalog}$ since it is the best performing
one and an approximation of the present search functionality.

\textbf{Entered by at least k Players}

We start off by presenting the results for the system $SE_{play,k}$ which indexes
all the tags that are entered by at least $k$ different players. As said before,
we varied the value of $k$ in the set $\{2, 3, 4\}$ and as the value of $k$ increased
the precision, recall, and MAP score monotonically decreased. The statistics

\textsuperscript{15} Again, we built the systems using the Xapian open source search engine library and
did not vary the parameters for the BM25 function but used the defaults provided by
Xapian.
### 4.6. Filtering Non-Topical Game Tags

#### Table 4.5: MAP/Precision/Recall scores for the search engines. Notational convention: the system that indexes the data obtained by applying the filter \( F \) is denoted by \( SE_F \); only the best performing filters for each filtering approach are shown. ↑, ↓, and ≈ indicate if a score is significantly better, worse, or statistically indistinguishable from the scores of \( SE_{\text{catalog}} \), \( SE_{\text{user}} \), and \( SE_{\text{vuser}} \), respectively.

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( SE_{\text{catalog}} )</td>
<td>0.168↑↑↑</td>
<td>0.438↑↑↑</td>
<td>0.291↑↑↑</td>
</tr>
<tr>
<td>( SE_{\text{user}} )</td>
<td>0.131↑↑↑</td>
<td>0.165↑↑↑</td>
<td>0.589↑↑↑</td>
</tr>
<tr>
<td>( SE_{\text{vuser}} )</td>
<td>0.081↑↑↑</td>
<td>0.193↑↑↑</td>
<td>0.286↑↑↑</td>
</tr>
<tr>
<td>Tag filters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( SE_{F_{\text{play,2}}} )</td>
<td>0.171↑↑↑</td>
<td>0.221↑↑↑</td>
<td>0.523↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{lda,80}}} )</td>
<td>0.149↑↑↑</td>
<td>0.222↑↑↑</td>
<td>0.454↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{prep,70}}} )</td>
<td>0.145↑↑↑</td>
<td>0.215↑↑↑</td>
<td>0.404↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{play,2+catalog}}} )</td>
<td>0.083↑↑↑</td>
<td>0.198↑↑↑</td>
<td>0.287↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{lda,80+catalog}}} )</td>
<td>0.199↑↑↑</td>
<td>0.233↑↑↑</td>
<td>0.603↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{lda,30+catalog}}} )</td>
<td>0.174↑↑↑</td>
<td>0.258↑↑↑</td>
<td>0.53↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{play,7+catalog}}} )</td>
<td>0.19↑↑↑</td>
<td>0.243↑↑↑</td>
<td>0.541↑↑↑</td>
</tr>
<tr>
<td>( SE_{F_{\text{play,2+catalog}}} )</td>
<td>0.15↑↑↑</td>
<td>0.247↑↑↑</td>
<td>0.449↑↑↑</td>
</tr>
</tbody>
</table>

For the best performing system \( SE_{F_{\text{play,2}}} \) are presented in Table 4.5. As we can see, \( SE_{F_{\text{play,2}}} \) performs only marginally better than the system that indexes the verified tags, \( SE_{\text{vuser}} \), and much worse compared to the system that indexes all tags, \( SE_{\text{user}} \). It seems \( F_{\text{play,2}} \) is able to detect only a small fraction of the good tags that are outside of the set of verified tags, hence the marginal increase in precision and recall. The combination of the tag filtering and catalog metadata worsens the matters. \( SE_{F_{\text{play,2+catalog}}} \) has higher recall than \( SE_{\text{catalog}} \) but comparatively lower precision which results in lower MAP score.

#### Latent Dirichlet allocation-based filtering

This section presents the results for the system \( S_{F_{\text{lda,30}}} \) which indexes only the top \( k \) tags from each fragment where the ranking is based on the a posteriori probability that a given tag is assigned to the fragment as estimated by the LDA model. The value of \( k \) was varied from 10 to 100 with an increment of 5. The best results are achieved for \( k = 30 \) and are shown in Table 4.5. As we can see, \( S_{F_{\text{lda,30}}} \) outperforms that system \( SE_{\text{user}} \) that indexes all user tags by 14%. However, the retrieval performance of \( S_{F_{\text{lda,30}}} \) falls short when compared with the systems \( SE_{\text{catalog}} \) and \( S_{F_{\text{lda,80}}} \) which outperform it by 28% and 31%, respectively. The combination of the catalog data and LDA tag filtering yields only a marginal improvement in performance: \( SE_{F_{\text{lda,30+catalog}}} \).
outperforms $SE_{F+catalog}$ only by 3% with respect to the MAP score.

**Discussion**  
A potential explanation for the poor performance of the LDA based filtering is the size of the corpus. Our tag/fragment collection is not large enough for the LDA model to provide a good estimation of the underlying topic structure. In fact, this was our suspicion all along however we decided to include LDA for the sake of completeness — since TF-IDF and LDA are among the most commonly used methods for measuring word to document relevance and topic inference.

**Player reputation-based filtering**

This section outlines the results for the system $S_{F_{prep,k}}$ which takes only the top $k$ tags for each fragment where the ranking is based on the combination of the TF-IDF score and the reputation of the players. As previously, the value of $k$ was varied from 10 to 100 with an increment of 5. The best results are achieved for $k = 70$ and are displayed in Table 4.5. As we can see there is an improvement of 11% with respect to the MAP score compared to the system $SE_{user}$ which indexes all user tags. However, the search performance of $S_{F_{prep,70}}$ is worse for all three metrics than the system $S_{F_{tfidf,80}}$ which ranks and filters the tags based only on their TF-IDF scores. This means that considering the reputations of the players in combination with the TF-IDF
4.6. FILTERING NON-TOPOCAL GAME TAGS

Figure 4.3: Search performance statistics for the systems $SE_{F_{tfidf,k}}$.

ranking in this case only worsened the matters for the TF-IDF based filtering. This conclusion is further supported by the fact that $SE_{tfidf,80+catalog}$ outperforms $SE_{prep,70+catalog}$ by 5% with respect to the MAP score.

Discussion Figure 4.2 shows the correlation between the reputation of the players and their success in providing tags that resulted in relevant search results. In particular, the horizontal axis shows the reputation for each of the users in increasing order whereas the vertical axis presents the number of relevant video fragments that can be retrieved using tags entered by the corresponding player. As seen, higher reputation does not entail higher success rate. In fact, many of the most reputable players are among the ones that were the least successful. This means that computing reputation in the manner outlined above may not be the best predictor for the ability of the player to provide topically relevant tags. Moreover, considering [92, 57] our reputation estimation method assigns higher reputation to players which settle on low-effort tags. Consequently, tags favored by this filtering method will tend to be more ‘obvious’. Alternative explanation is that there is not sufficient data available to reliably estimate the reputation of the players. Recall that 99% of the players are registered as anonymous and the only data available for them is a single session of playing Waisda?

TF-IDF-rank Take Top k

This section outlines the results for the systems $S_{F_{tfidf,k}}$ which consider only the top $k$ tags for each fragment where the ranking is based on the TF-IDF
scores of the tags. We vary the value of $k$ from 10 to 100 with increments of 5. Figure 4.3 presents the search performance statistics as the value of $k$ varies. Not surprisingly, as the value of $k$ increases the search precision and the search recall decrease and increase, respectively. Moreover, the MAP score steadily increases until $k$ equals 80, where the maximal value is reached, and then it starts to decrease. As shown in Table 4.5, the MAP score of $S_{tfidf,80}$ is 0.171 which means it significantly outperforms $SE_{user}$, the systems that indexes all user tags. Indeed, the increase in search performance is 31%. What is more important is that $S_{tfidf,80}$ slightly outperforms even $SE_{catalog}$, which until now was the best performing one and the system that we are trying to beat. This suggests that TF-IDF score is indeed a good indication of the quality of the tags as topical descriptors and can be used to filter out the non-topical tags. Furthermore, the combination of catalog metadata and TF-IDF filtering proves to be beneficial: $SE_{tfidf,80+catalog}$ outperforms $SE_{catalog}$ by 18%. The improvement is caused by the fact that the TF-IDF ranked tags increase the search recall of the catalog metadata by factor of 2; the recall of $SE_{tfidf,80+catalog}$ is higher than the recall of $SE_{catalog}$ by 107%.

4.7 Conclusion and Discussion

In this chapter we studied to what extend players enter tags which are valid topical descriptors of the video material. Another aim was to derive insights about players’ tagging practices that are associated with ascribing topical tags. The study was carried out with the focus on topical search.

In Section 4.5 we evaluated the search performance of the entire unprocessed collection of the user tags. The general conclusion is that the search performance of the raw, unprocessed, user tags for retrieving videos based on topic leaves much to be desired. While the search recall of the user tags is relatively satisfactory the precision is rather poor. Our analysis showed that a significant portion of the topics are indeed captured by the user tags. However, there are also many user tags that do not pertain to the subject (semantics) of the video, but refer to the more syntactic aspects such as what is seen or heard. It is the latter group that is responsible for the false positives and thereby hurting the search precision. Therefore, if user tags are to be used for topical search, a preprocessing step is required that will identify and filter out the non-topical user tags.

The quality of the tags as topical descriptions was addressed in Section 4.6 where we looked into several ways we can detect and filter out the non-topical user tags. While the different methods that we studied performed with various success, the conclusion is that the game tags can be successfully
exploited for topical video search provided there is a filtering process that would reduce or eliminate the effect of the non-topical tags. Our results show that after TF-IDF-based filtering game tags can emulate the retrieval performance of the best performing system that utilizes manually crafted metadata for search. Moreover, combining TF-IDF filtered game tags with the manually crafted metadata yields an improvement of retrieval performance by 18%. The improvement is attributed to the increased retrieval recall stemming from the game tags. An important consequence of this result is that tagging games provide a cost-effective alternative for AV collection owners that do not possess the required manpower to manually annotate their material.

Successfully deploying a GWAP is no small feat. Attracting new players and keeping them engaged over time is vital for success and requires continuous publicising efforts. The experience gained from the Waisda? project showed that targeting the fanbase of the TV series being tagged in the game is an effective method of attracting new players. Player’s engagement can be sustained with in-game motivational mechanisms such as leaderboards, and with time-limited contests offering awards for the top performers. Planning and executing such activities is costly, however the cost is independent of the size of the video collection. The cost of manual annotation, on the other hand, increases linearly with the size of the collection and will eventually outweigh the Waisda?-related costs.

What makes video annotation a difficult task is the fact that video is a medium that is extremely rich in meaning. The taggers can get overwhelmed by the complex interplay of objects and events especially in the fast-paced game setting. Our qualitative analysis showed that significant portion of the Waisda? tags refer to more noticeable aspects of the content such as moving objects in the background or in the foreground, or prominent stationary objects. This is hardly surprising, after all these are things that are easy to reach consensus on and ultimately that is the goal of the game from the perspective of the players. We also observed that the tags which are present in the audio and refer to a concept that appears visually are more likely to be topical descriptors. Their presence in both the audio and visual component is an indication that they are denoting salient aspects of the content. Therefore, if the goal of the AV collection owners is collecting topical tags, this insight can be operationalized in the game by instructing the players to tag things that are both on screen and in the audio.

Waisda? is a production grade open-source crowd-sourcing tool which is relatively easy to set-up and after the suitable processing the collected tags can be exploited for retrieval. We believe that this is an important step toward making the AV heritage more accessible on the Web and in general
making the Web more connected.
Chapter 5

Waisda?: Open-source Video Labelling Game

In this chapter we shall describe in more detail the video labelling game Waisda? and how it can be exploited as a crowdsourcing tool for collecting user-generated tags for video fragments. We shall also outline the results and efforts from this thesis that had an impact on the development of Waisda?.

This chapter is based on the paper entitled Waisda?: video labeling game which was presented at Proceedings of the 21st ACM international conference on Multimedia, held in Barcelona, Spain. The paper is authored by Michiel Hildebrand, Maarten Brinkerink, Riste Gligorov, Martijn Van Steenbergen, Johan Huijkman, and Johan Oomen.

5.1 Introduction

In essence, Waisda? is a Game With A Purpose (GWAP). GWAPs are a human-based computation technique in which a computational process performs its function by outsourcing certain steps to humans in an entertaining way [2, 122]. The crux of this approach is the realization that humans are better than computers in certain tasks. For example the computers still lack the basic conceptual intelligence or perceptual capabilities required for carrying out the task of labelling random images and videos [2]. In fact, the first example of a GWAP, the ESP Game created by Luis von Ahn, labels images on the Web by harnessing the capabilities of the players. The ESP game randomly pairs up two players with the task to describe images. Players don’t know who their partner is, nor can they communicate with each other. The
only thing partners have in common is an image they can both see. When both players provide the same label for an image, they score points and proceed to the next image. The labels entered by both users are associated to the image as metadata. In other words, the consensus among players is a mechanism to ensure the quality and consistency of the labels. The same design principles are applied in Waisda? as well. Waisda? is a multi-player labelling game where players compete with each other by tagging streaming video. The players score points by entering the same tag as one of the other players within a ten seconds interval. As a result, the video that is tagged in the game is annotated with the matched tags which are anchored to the time points in the video when they were entered.

Waisda? was conceived by the Netherlands Institute for Sound and Vision (S&V) and the VU University Amsterdam in the context of the Dutch project Images for the Future\(^1\) and the European research project Presto-PRIME, as an open-source crowdsourcing tool that can be deployed by multimedia collection owners as crowdsourcing initiatives to annotate their collection items. The development of the software was contracted to Q42, a Dutch internet development company\(^2\). The Netherlands Institute for Sound and Vision in cooperation with Dutch broadcasters deployed Waisda? in two consecutive pilots featuring historical archive material as well as recent TV episodes from Dutch shows. The first pilot of the game run from May 2009 until January 2010 and produced over 420,000 user tags. The follow up analyses of the collected tags \([118, 37]\) and user evaluations suggested potential areas of improvement. Considering the acquired experience and insights, S & V launched a second pilot of the game\(^3\) which featured several improvements, albeit the basic idea of the temporal tag agreement is preserved. The two pilots together resulted in more than a million tags describing thousands of videos. In fact, the studies described in Chapter 2 and 3 are based on the data collected in the first and the second pilot, respectively.

The open source version of Waisda? enables maintainers of an online video collection to start their own crowdsourcing initiative for time-based annotations. In particular, we believe that Waisda? is a valuable tool for cultural heritage institutions, as it provides a novel and engaging way for the public to access and interact with the audiovisual material.

In the rest of this chapter we shall describe the gameplay and the user interface of Waisda?, explain the software architecture, and give examples of successful deployment of Waisda? by third parties. In the final section

\(^1\)http://imagesforthefuture.com/
\(^2\)http://q42.com/
\(^3\)At the time of writing, the game is deployed online and can be accessed at http://woordentikkertje.manbijthond.nl/.
we outline the results and efforts from this thesis that had an impact on the development of Waisda? We conclude this section with a list of the awards\footnote{The list of awards is complete up to the time of writing.} given to the Waisda? project

- Waisda? won the Best Archives on the Web award in the category Best Use of Crowdsourcing in 2010.
- Waisda? won the TIMAF European Best Practices award in 2009

5.2 Gameplay and User Interface

The gameplay of Waisda? is straightforward: the player first selects a video from the Homepage, plays the Game by watching the video and entering tags, and finally studies the results in the Game recap to learn what he/she can improve in future games.

5.2.1 Homepage

Figure 5.1 shows the homepage of Waisda?, with six videos that the user can choose from. A player starts a game by selecting one of these videos. To give other players the opportunity to join the game, the player first enters the waiting room. In the waiting room the players can study the game instructions and after 20 seconds the game is automatically started. On the homepage the games that are waiting for players are shown at the bottom right. Players can quickly join a game by selecting one of the entries. The homepage also contains the basic instructions that prepare the user for the game. To honour the top players, and motivate the other players, the homepage contains a leaderboard that displays the top players of the week. To engage the players with the purpose of the game a notification at the top of the page shows the total number of tags that were entered into the system, and the users contribution to this total. This part of the page also contains the links to register and login. Games can be played without registering, but only for registered players the score is maintained in the player’s profile. To explain the potential of the contributed tags, the page contains a tag cloud with the most popular tags. By selecting a tag the user can navigate the video collection. Finally, the bottom of the page has space reserved for logos and navigation links to additional pages, such as an about page or a page with the terms of use.
5.2.2 Game

Figure 5.2 shows the game page of Waisda?. It contains a video player and below it a text-entry field. When the player enters the game page the video automatically starts playing, the text-entry field receives focus, and the player can start entering tags. The right side of the page contains the scoreboard. It consists of the current score of the player, the current rank and a listing with all tags entered by the player. The tags are displayed with their score and an icon indicating the type of match and the tag type (person, location etc.).

Waisda? uses a scoring mechanism to motivate users to (keep) playing the game, and reward them for entering specific types of tags. The basic scoring mechanism is tag agreement, two players entering the same tag. From the perspective of the video collection tag agreement provides a mechanism to improve the quality, tags that are entered by more than one user are less
likely to contain spelling errors and are more likely to provide a trustworthy description of the video. Tags are considered a match if they are syntactically the same after normalization and entered within a ten second interval of each other. The details of the normalization procedure are explained in the documentation. The matching procedure can be extended by importing tag similarity lists. For example, tags can be matched semantically by importing synonym lists, and specific tags (e.g. Labrador) can be matched with more generic tags (e.g. Dog) by importing word pairs derived from the hierarchical structures of linguistic or domain specific thesauri.

Waisda? also contains a mechanism to reward players for entering specific types of tags. For example, when tagging videos in the art domain the names of artists are important. By importing a dictionary with artist names Waisda? detects these tag types and rewards the players additional points.

As we can not expect that there are always sufficient users active at the same time, the scoring mechanism of Waisda? also motivates users to play alone. In this case a player scores points by matching with the tags that were entered for the same video in previous games. To ensure players not only tag in the most obvious ways, the scoring mechanism of Waisda? also motivates players to pioneer new tags. A player enters a pioneer tag when this tag is not entered for this video before, and afterwards it is matched with tags entered by other players (in the same or a later game).

5.2.3 Game Recap

When the video ends the user is automatically redirected to the game recap (see Figure 5.3). On this page the user can investigate his/her performance, learn about the details of the scoring mechanism and the behaviour of other players. The game recap consists of two parts. The tag statistics, placed on the left part of the screen, give an overview of the number of tags entered by the player in different categories, including the number of matched tags, pioneer tags and specific types of tags. The tag details, placed on the right part of the screen, are listing of all the tags entered by the player. The listing is similar as the one provided in the game, but includes information about other player in the game, if any, who entered the same tag within a ten seconds interval i.e. a tag match has been achieved.

5.3 Implementation

Waisda? is an open source crowdsourcing tool that has been designed and implemented as a 3-tier web application. The backend application logic is
Figure 5.2: Screenshot of the Waisda? game page with public domain video clip from Preligner Archives.

Figure 5.3: Screenshot of the Waisda? game recap after a successfully finished game.
5.3. IMPLEMENTATION

implemented in the Spring Framework\(^5\), which at the time of writing is one of the most widely used open source Java application frameworks. The application is backed by a MySQL\(^6\) database which is accessed via the Java Persistence API and Hibernate\(^7\). The user interface (UI) is rendered using the Java Server Page (JSP) technology. The UI consists of HTML, JavaScript, and CSS. jQuery and Twitter Bootstrap are used as the JavaScript and CSS libraries. The dynamic stylesheet language LESS is used to allow easy configuration of the presentation style.

The Waisda? source code and documentation is available from Github \(\text{http://github.com/beeldengeluid/waisda}\). The building, installing and running or the Waisda? source code is handled by Maven\(^8\) which is a open-source software project management tool developed by the Apache Foundation. The getting started section of the documentation contains instructions to setup a deployment using videos from the Prelinger archive\(^9\).

5.3.1 Backend

After installation the maintainer has to populate the database with videos and optionally with dictionaries and similarity lists. A video has a title, description, source URL, key frame URL, duration and video player type. There are currently two types of video players supported: the JWPlayer (HTML5 and Flash), and the NPO player (Silverlight). The latter is used to play content from Dutch public broadcasting associations. Each video in the project specifies which of the two players it would like to use, in combination with the parameters required for the players. To facilitate more robust scoring schemes auxiliary resources like dictionaries and similarity lists can also be used. As long as the data format is valid, the party that deploys the game is free to use any resources it pleases. A dictionary is a set of word-type pairs which is used to reward additional points for specific types of tags. Examples are names of celebrities and names of geographical locations. A similarity list is a set of word-word pairs which is used to extent the default syntactic tag matching with semantic similarity matching. Examples are synonym pairs or words that are hierarchically related (specific-generic).

\(^{5}\)http://spring.io/
\(^{6}\)http://www.mysql.com/
\(^{7}\)http://hibernate.org/
\(^{8}\)http://maven.apache.org/
\(^{9}\)http://archive.org/details/prelinger
CHAPTER 5. Waisda?: OPEN-SOURCE VIDEO LABELLING GAME

Figure 5.4: Screenshot of the default user interface of Waisda? stripped of all branding and styling.

5.3.2 Frontend

The default user interface of Waisda? is completely functional. In addition, basic branding and styling can be done by adding a logo and changing the color scheme. These changes can be made with little knowledge of the system and are described in detail in the documentation. The documentation also describes how a maintainer can modify the advanced functionality of the system, such as the videos that are shown on the homepage, how tags are matched and which scores are given. It also describes how new pages can be added or how the page layout can be changed. Making such changes requires more advanced knowledge of HTML, CSS and Java Server Pages. Figure 5.4 shows the basic Waisda? interface stripped from all branding and styling whereas Figure 5.5 shows an experimental setup of Waisda? styled with the logo and the colour scheme of the European Film Gateway\(^\text{10}\).

\(^{10}\text{http://www.europeanfilmgateway.eu/}\)
5.4 Waisda? Deployments

In this section we present several examples of deployments of Waisda? in the wild.

5.4.1 Europeana

Perhaps the most notable deployment of Waisda? was in the context of the Europeana project\(^\text{11}\) where Waisda? was included as a showcase\(^\text{12}\). Europeana is an internet portal that provides multi-lingual access to millions of books, paintings, films, museum objects and archival records that have been digitised throughout Europe. More than 2,000 institutions across Eu-

---

\(^\text{11}\)http://www.europeana.eu/
\(^\text{12}\)http://labs.europeana.eu/apps/waisda-floss
rope have contributed to Europeana. These range from major international
names like the Rijksmuseum in Amsterdam\footnote{https://www.rijksmuseum.nl/}, the British Library\footnote{http://www.bl.uk/} and the Louvre\footnote{http://www.louvre.fr/} to regional archives and local museums from every member of the European Union. Together, their assembled collections let users explore Eu-
rope’s cultural and scientific heritage from prehistory to the modern day.
The Europeana project incorporates the tools, solutions, and preservation
practices developed in the PrestoPRIME project\footnote{As mentioned already Waisda? is one of the deliverables of the PrestoPRIME project. The publicly available deliverable report can be found at [47].}. One of the main contrib-
utions of the PrestoPRIME project incorporated in Europeana is an OAIS-
compliant preservation framework capable of supporting the whole range of
digital preservation operations in the long term. Waisda? can be integrated
in the workflows and processes included in this framework that deal with
user-generated annotations for videos. In order to provide programmable in-
teraction between Waisda? and the other systems in the framework, in the
scope of this thesis, an extension module of Waisda? was developed. The
module consists of a number of web services that enable adding/updating
videos that are to be tagged in the game and extracting the collected tags in
a light-weight and highly-portable textual format. More details on this can
be found in section 5.5.1.

5.4.2 Spotvogel

Another application of Waisda? was the so-called Spotvogel\footnote{http://spotvogel.vroegevogels.vara.nl/} which translates to mocking bird from Dutch. Spotvogel was deployed by the Dutch broadcaster VARA\footnote{http://www.vara.nl/} and the S&V institute with the aim of collecting user
tags for the footage from the Dutch television program Vroege Vogels\footnote{http://spotvogel.vroegevogels.vara.nl/} (Early Birds in Dutch) which is about wildlife. The game run for 6 months in 2013
and was nominated for Dutch Game Awards 2013\footnote{http://www.dutchgameawards.nl/2013/spotvogel/ note that the link is in Dutch}. The goal of the game
from the perspective of the players was to identify and tag occurrences of
birds and other wildlife in the footage. The players were encouraged to be
as specific as possible when identifying the species of the wildlife they were
tagging. For example, a tag such as “long-eared bat” yielded more points
than the more generic tag “bat”. This was achieved by using controlled vo-
cabulary which contained various names of bird species. Whenever a player
entered a recognizable bird species name she or he was awarded extra points. This positive reinforcement was meant to stimulate and reinforce this kind of tagging behaviour in players. The broadcaster VARA counts on the collected tags to increase the accessibility and the searchability of their archive.

5.4.3 European Film Gateway Experiment

This example differs from the previous ones in the fact that in this case Waisda? was used as a tool to carry out an experiment. The experiment was designed and performed by Liliana Melgar Estrada during her stay at the VU University Amsterdam and resulting paper [30] was accepted for publication in the Journal of the Association for Information Science and Technology. The experiment focuses on studying the differences between how film experts and novices describe (tag) fiction films. As a source material, five videos (fiction clips) from the European Film Gateway were chosen. The act of annotating the videos by the experts and novices was carried out using a customised Waisda? installation set up specifically for this purpose. What is worth mentioning is that the customisation capabilities available in Waisda? out of the box were sufficient to satisfy the requirements put forward by the experimental design.

5.5 Thesis Contributions

In the final section we outline the contributions of this thesis for the development of Waisda? which fall into two areas software development and game design.

5.5.1 Software Development

As mentioned in Section 5.4.1, Waisda? can be integrated into the OAIS-compliant preservation framework developed in the PrestoPRIME. In order to provide programmable interaction between Waisda? and the other systems in the framework, we developed an extension module of Waisda?. The module was used in the PrestoPRIME technical testbed held in Turin 2011\textsuperscript{21}, with the explicit purpose of ingesting tags collected with Waisda? into the preservation framework. The module is now part of the open-source Waisda? codebase available on GitHub\textsuperscript{22}.

\textsuperscript{21}https://www.prestocentre.org/events/prestoprime-technical-testbed-event
\textsuperscript{22}https://github.com/beeldengeluid/waisda
The module has been developed in Java programming language and uses the same technology stack as the Waisda? core. In particular, the application logic is implemented in the open-source Spring framework and the access to the database is facilitated by the Java Persistence API and Hibernate. Figure 5.6 shows the architecture overview and the flow of data among the Waisda? core, the module, and external systems. The Waisda? core and the module communicate through shared database. Both components read and write data to the same underlying MySQL database. In literature, this is referred to as shared database enterprise integration pattern [50]. All direct interaction from the outside world with the database goes through the module which acts as a gateway. External systems read and write data using the services defined by the module. External systems and the module exchange messages encoded in the lightweight highly-portable data-interchange JSON\textsuperscript{23} format. This allows for high-level of decoupling and interoperability between Waisda? and external systems which can be implemented in any programming language without any knowledge about the underlaying technical details of Waisda?.

The module consists of the following services:

- **Video submission service.** Using this service external systems can add new videos for tagging in Waisda?. Figure 5.7 shows the format of the JSON message that is expected by this service. As seen, the external system needs to provide basic video metadata, such as the title, duration, URL from which the video can be streamed, fragment ID used

\textsuperscript{23}http://www.json.org/
5.5. THESIS CONTRIBUTIONS

Figure 5.7: JSON data request payload for the video submission/update web service. The data payload consists of metadata for the video that is to be submitted or updated.

```json
{
    "title": "American flamingo",
    "playerType": "JW",
    "sourceUrl": "http://goo.gl/Jw577N",
    "fragmentID": 1,
    "imageUrl": "https://goo.gl/NmBq47",
    "enabled": true,
    "startTime": 0,
    "duration": 740000
}
```

Figure 5.8: JSON data request payload for the enable videos for tagging web service. The data payload consists of an object which contains a fragmentId which identifies the video to be acted on and a boolean enabled flag.

```json
{
    "fragmentID": 1,
    "enabled": true,
}
```
Figure 5.9: JSON data response payload for the export tags for video web service. The data payload consists of an array of tags entered after the specified timestamp.

to identify the video subsequently, video player type, etc. It should be noted that the video itself is not uploaded to Waisda?, instead the video is streamed from the provided URL. When calling this service it is assumed that no other video with the provided fragment ID exists already. Failing to comply with this requirement will result in an error.

• **Video update service.** Using this service the metadata of videos already within Waisda? can be changed. This service accepts the same type of JSON message as the submission service above. It is assumed that video with the specified fragment ID exists already. If this condition is not satisfied, an error will be raised.

• **Video enable/disable service.** Using this service external systems can enable/disable a video for tagging in Waisda?. Figure 5.8 shows the format of the JSON message that is expected by the service. The fragment ID of an already existing video in the system needs to be provided along with the flag that determines whether the video should be enabled or disabled.

• **Video tags export service.** Using this service all tags for a specific video can be exported in JSON format. The data being exported can be
limited to tag entries after a certain timestamp, enabling efficient sync mechanisms. External systems need to provide a fragment ID of an existing video in the system and an optional time-stamp. If a time-stamp is provided, all tags for the video entered after the specified time-stamp will be returned. Otherwise, all tags for the video will be returned. Caller external systems are advised to provide a time-stamp to reduce the number of tags returned. This is merely an optimization step. Figure 5.9 shows the format of the result returned by this service.

It should be noted that, while initially developed for the Europeana environment, these services are not bound to it in any way and can be seamlessly integrated in any system or workflow. The only requirement is that the external systems ought to honor the contract defined by the services when calling them. The services provide a basic set of operations that can be used to implement more complex real-life scenarios involving Waisda? as a crowdsourcing tool for collecting video annotations.

5.5.2 Game Design

The development of Waisda? proceeded in two iterations. After the first pilot was launched the follow up analyses of the collected tags [118, 37] and user evaluations suggested potential areas of improvement. The suggestions were incorporated into the game design and the second, which is also the current, version of Waisda? was rolled out.

One of the follow up studies [37] was carried out in the context of this thesis and it is covered in detail in Chapter 3. In this study we noticed limitations or the game tags collected in the first pilot. One limitation is the low number of specific type of tags in the who and where facets i.e. persons and locations. We showed that by matching the tags to controlled vocabularies we can derive the type of the tags (person, location, organization, etc.). We suggested a modification in the scoring mechanism where the players are awarded extra points for entering specific types of tags e.g. persons or locations (see Section 5.2.2 for details). This positive reinforcement is meant to stimulate and reinforce this kind of tagging behaviour in players. Another limitation of the first version of Waisda? was the way the tags entered by different players were compared, which was basically a case-insensitive string comparison. We suggested extension of the matching mechanism where synonyms, hyponyms, and hyperonyms would also be considered a match (see Section 5.2.2 for details). Both suggestions were approved and are now part of Waisda?.
The other follow up study [118] was conducted by Annelies van Ees in the context of her master thesis project. The aim of the study was to determine factors that motivate people to play a video tagging game such as Waisda?. The conclusion of the study was that players are more inclined to play when videos are shorter and more feedback is provided. The feedback in this case was in the form or game recap once the game is finished. Based on these findings the second pilot included short videos and also featured a game recap screen (see Figure 5.3 for details).

5.6 Conclusion

In this chapter we described Waisda which is an open-source crowdsourcing tool that can be deployed by audio-visual collection owners to annotate their collection items. We outlined the main technological aspects of the game, the gameplay, and the key user interface features. At the time of writing, the latest release of Waisda is a mature, production-grade, customizable game with a purpose, with a diverse usage. Many aspects of the game can be customized, such as the user interface and tag matching mechanisms, in order to fit various application scenarios.

As part of the research described in this thesis, an extension module to Waisda? was developed which enables seamless integration of the game in broader audio-visual workflows. The module features a set of high-level services which hide the underlying technological complexity and provide communication in the light-weight and highly-portable JSON textual format. The services offer support for the basic add, update, read, and enable/disable operations which serve as a basis on which more complex integration scenarios can be built. Having said that, and considering the results from Chapters 3 and 4, Waisda? is valuable tool at the disposal of the audio-visual collection owners which provides cost-effective way of obtaining community-centric video annotations.
Chapter 6
Conclusions and Discussion

In this chapter we revisit the research questions outlined in Chapter 1 and state the conclusions that follow from the research work presented in this thesis. We conclude with a final discussion and indicate possible research directions to extend or compliment the results described in the thesis.

6.1 Main Contributions

As audio-visual archives become accessible to the wider internet audience, successful retrieval of archive items is an increasingly challenging task. One of the most important reasons for this is the lack of adequate annotations. Users cannot find what they are looking for because annotations are either not present or too expert-centric. Video tagging games are an attempt to alleviate this problem. Engaging the internet community to tag the videos addresses the issue of scarcity of annotations and yields tags that are community-centric. The focus of this thesis is on the role of game tags in audio-visual archives and their added value for video retrieval. To understand what do game tags bring to the table as opposed to the catalogers’ annotations we study the differences between them along two dimensions. The first dimension is the terminology: we perform a quantitative study to estimate the terminological gap between taggers and the catalogers. The second dimension is the descriptive scope i.e. which aspects of the video content are described and at what level of abstraction and granularity. We carry out a qualitative study in which we analyse the descriptive scope of the game tags. Making the videos more accessible to the wider internet audience is one of the main objectives of the
game tags. To evaluate the efficiency of the game tags for video retrieval we design Cranfield-style experiments for two prominent scenarios that arise in practice: visual and topical search. Moreover we perform a qualitative analysis of samples of search results to establish the reasons that usually lead to true and false positives in both scenarios. The fast pace of Waisda? usually results in players contributing mainly non-topical tags. To combat this effect in the topical search scenario, we characterize the quality of user tags as topical annotations and identify several tag features (filters) that serve as indicators whether game tags are useful for retrieval. The success of each of the filters is evaluated in Cranfield-style experiment.

The main contributions of the thesis are:

- A method to analyse quantitatively the overlap between user and professional terminology exploited in video annotation.

- Qualitative analysis of the aspects of the video content which are described by the game tags i.e. *what do game tags describe in a video and how.*

- A method to design a dataset —including fragments, queries, and relevance judgements — for evaluating the added value of user tags for visual and topical video search.

- Two evaluation datasets —one for visual and one for topical search — which contain fragments, user tags from Waisda? collection, queries, and relevance judgements.

- Quantitative evaluation of the added value of the game tags in terms of visual and topical video search.

- A method for qualitative analysis/classification of the search results in the context of visual search. The aim of the analysis is to discover the types of tags that are generally responsible for false positives.

- Qualitative analysis of the search results in the context of topical search. The aim of the analysis is to discover tagging practices that generally produce relevant topical tags.

- Evaluation of tag-related measures for accessing the quality of the game tags as topical descriptors in the context of topical video search.
6.2. RESEARCH QUESTIONS REVISITED

- Development of extension module for Waisda? which enables integration of the game in AV collection workflows. The module is part of the latest release of Waisda? which is available online\(^1\).

In this chapter we consider again the research questions stated in Chapter 1 outlining the methods and results. The last section features discussion and directions for future research.

6.2 Research Questions Revisited

6.2.1 What is the relationship between the game tags and professional annotations in terms of vocabulary and what they describe?

In the typical audio-visual archive ecosystem the game tags coexist with professionally curated annotations. Successful integration of the game tags into the AV collections’ workflows requires a better understanding of the characteristics of the tags compared to the professional annotations. To this end, in Chapter 2, we perform a quantitative and a qualitative study of the game tags in which we compare them to the professional annotations along two dimensions: (i) the vocabulary used i.e. the terminological gap and (ii) the descriptive scope i.e. which aspects of the video content are described and at what level of abstraction and granularity.

Studying the Terminological Gap

In order to study the terminological gap we devise and carry out a quantitative analysis of the game tags. Presently, Waisda? has no built-in mechanisms that control the quality of the language used by the taggers. Therefore, the first goal of the analysis is to determine the fraction of the game tags that are meaningful words. For this purpose we exploit two vast background knowledge sources — the general lexicon of the Dutch language Corneto and the Google search engine — as semantic filters; whenever a tag is present in either of these two sources we take it as a signal that the tag is a meaningful word. Additionally, Corneto encodes part-of-speech information about the words which we exploit to quantify the types of words generally used by the taggers. Corneto also provides information about the various senses of the words which is used to detect polysemy. The results from the

\(^1\)The Waisda? source code and documentation is available from Github [http://github.com/beeldengeluid/waisda](http://github.com/beeldengeluid/waisda)
analysis show that the majority of the game tags are meaningful words in themselves. This analysis also reaffirmed some of the inherent limitations of the user tags compared to the controlled vocabularies; namely, subjectivity and ambiguity of the language. Using the part-of-speech information we detected that the majority of the tags found in Cornetto are nouns and verbs. This means that people generally focus on objects and actions when tagging the videos. Small fraction of the game tags found in Cornetto, around 15%, are adjectives. Generally these are subjective descriptive words, for example the most common adjective among the tags is lekker (nice in Dutch). The general concern is that the overly subjective and imprecise tags can hinder searching. We come back to this in Section 6.2.2.

Finally, to check whether there is a terminological gap between the game tags and professional annotations we matched all game tags to the domain thesaurus GTAA. GTAA is the sole vocabulary that is employed by the catalogers when annotating the same videos that are tagged in the game. We observe that only a small fraction of the unique game tags can be found in GTAA, around 8%. This showed that there is indeed a terminological gap in the way the gamers and catalogers describe the videos.

The Descriptive Scope of Game Tags

To get a better understanding of what aspects of the video content are described by the game tags we carried out a manual classification of a sample of game tags. We used a classification schema that combines the Panofsky-Shatford model and a classification framework built by Hollink et al [52]. The results of the tag classification demonstrate that the game tags predominantly describe objects using generic language and rarely refer to topics of a scene or the entire video. This is in sharp contrast with the professional annotations which exclusively target the topic(s) of the entire video and (less frequently) of particular scenes. The general conclusion is that game tags and the professional annotations are complementary both in terms of the terminology and in terms of what aspects of the video content they describe. This makes the game tags a valuable asset for at least the following two reasons. First, they provide a more online community-centric view of the content which will undoubtedly improve the access for the wider internet audience. Second, game tags are fine-grained annotations that target scenes, events, and objects within scenes. At this level of granularity they can facilitate tasks like event detection, instance search etc. Moreover, this level of granularity generally is unattainable for the professional annotations due to the enormous effort and time required. Game tags can step in and fill the void.
6.2. RESEARCH QUESTIONS REVISITED

6.2.2 Can we improve video retrieval with the help of user-generated data?

Video fragment retrieval is widely recognized as an application scenario of particular business importance in the AV archival world. We evaluate the effectiveness of the game tags for two retrieval tasks: (i) visual search, retrieval of fragments that feature visual appearances of the concept of interest and (ii) topical search, retrieval of fragments that are about a given topic.

Game tags for visual search

When evaluating the effectiveness of game tags for visual search in Chapter 3, we are mainly interested in two research questions. First, can game tags on their own or in combination with other existing types of metadata improve search? And second, does using only verified (consensus) tags give better search performance than using all game tags? In addressing these questions we use Cranfield-style quantitative evaluation methodology. This methodology encompasses creation of an evaluation dataset that consists of a query set and relevance judgements. The query set is derived from real-life query logs. The relevance judgements are at the hearth of this methodology and reflect the fact that queries ought to be interpreted visually. We devised and carried out a crowdsourcing experiment to gather them by harnessing the efforts of the online Waisda? community.

The results from the evaluation show that game tags are doing exceedingly well in visual search. They outperform the other types of metadata in pairwise comparison. Moreover, combining the game tags with the other metadata types proved to be beneficial for visual search as well. In fact, including the game tags boosts the search performance for 33%. On the other hand, considering only verified tags results in worse search performance than considering all game tags. This is mainly caused by the fact that search based on all game tags yielded more relevant results and resulted in significantly higher recall. It seems the tag verification criterion is too conservative in a sense that it filters out tags that are in fact useful for search.

The overall conclusion is that game tags are well-suited for retrieving fragments that visually depict given object of interest. In Chapter 3 we establish that game tags usually describe events and objects within scenes so it is not surprising that they excel at visual search. We also establish that there is a difference in the focus and the scope of the game tags compared to the professional annotations; game tags being more fine-grained and professional annotations being more coarse-grained referring to the entire video and describing the prevalent topics. This property of the game tags makes
them particularly appealing for business use-cases such as discovering/selling stock footage. Moreover, recent advancements in areas such as computer vision open an array of possibilities for hybrid approaches. This is discussed more in Section 6.3.1.

**Game tags for topical search**

In Chapter 4 we study the added value of game tags for topical search. To this end, we use once again the Cranfield-style quantitative evaluation methodology. Central to this methodology is the creation of relevance judgements. For the purpose of the study, we devise a procedure that automatized the process by exploiting the in-house catalog tags as the gold standard w.r.t. what topics are covered in the fragments. This allows us to consider much larger set of fragments since we are no longer dependent on human judges.

The collected retrieval performance metrics show that when it comes to topical search the game tags leave much to be desired. While the retrieval recall of the game tags is relatively satisfactory the precision is rather poor. In order to understand the causes for the poor retrieval performance we devise and carry out a qualitative analysis of samples of true and false positives retrieved using the game tags. The analysis encompasses manual classification where the tags in the samples are classified w.r.t. to the content component they refer to — audio, visual, or both — and their descriptive scope (see Section 4.5.2 for more details). The aim of the analysis is to investigate the relationship between the content component and the descriptive scope of the game tags on one side and whether they resulted in true of false positive on the other.

The results of the analysis demonstrate that the overwhelming majority of the false positives are caused by game tags that do not pertain to the subject (semantics) of the video, but refer to the more syntactic aspects such as what is seen or heard. For example, many of these cases refer to general tags describing objects appearing in the foreground and the background of the scenery. It seems that the fast-pace of the game and the scoring mechanism which rewards tag consensus encourages the players to focus more on the immediate short-lived aspects of the video content rather than the more high-level topical aspects. After all it is easier to reach an agreement on prominent objects that draw the attention than on scene topics. The current game design of Waisda? creates a bias that works against collecting topical tags which is evident from the difference in the results between visual and the topical search studies.

What is more important, our analysis showed that a significant portion of the topics of the videos are indeed captured by the game tags. The topical
tags are present albeit buried under the myriad of non-topical tags. Therefore, if the game tags are to be used for topical search, a preprocessing step is required that will identify and filter out the non-topical game tags. We addressed this issue in Chapter 4 and we come back to it in the Section 6.2.3.

**Shortcoming of game tags for retrieval**

One of the main drawbacks of the user-generated tags compared to the controlled vocabularies is that tags are inherently imprecise. The players of Waisda? are unconstrained when entering tags and this can lead to several limitations. There is no explicit meaning assigned to the tags which can result in ambiguity. The prototypical example is the word *bank* which may refer to a river *bank* or to a financial institution. Without explicit disambiguation there is no way to tell the intended meaning. Another limitation of game tags stems from the presence of *synonyms* in natural languages. For example, both the tags *mac* and *macintosh* can be used to denote Apple Macintosh computer. If the searcher searches for *mac* but only *macintosh* is present in the system the search will be unsuccessful. Yet another limitation is imposed by the general level of *tag literacy*. Tag literacy in this sense refers to the ‘etiquette’ of generating tags in a way that increases their social value, balancing individual needs with the needs of the community. Low level of tag literacy stems from the lack of universally accepted tagging conventions among taggers. This may lead to the usage of specialised tags and “nonsense” tags designed as unique markers that are shared between a group of friends or co-workers. These limitations are not exclusive to game tags. The other metadata types suffer from them as well, albeit to a lesser extent. Closed-captions and in-house catalog tags can be ambiguous and have issues with synonyms as well. Tag literacy is a less of a concern for in-house catalog tags given their more controlled nature.

In the search studies described in Chapter 3 and 4 we use standard information retrieval mechanisms to combat the tag limitations outlined above such as stemming which is a standard technique to deal with word inflexions by reducing the inflected words to their root(stem). Tag literacy was not an issue in our search studies as the queries were all standard words and the word inflexions were taken care of by stemming. In other words, the effect of low literacy tags, was not triggered by the queries in our query set. Ambiguity was also not an issue since the majority of the queries in our query set are not polysemic. And even when there is more than one meaning assigned to the queries usually the difference among them is nuanced so they can effectively be used interchangeably. Synonymy can be addressed by query expansion where the synonyms are added to the query and searched for as
well. It will be interesting to extend the studies from Chapter 3 and 4 to deal with synonyms in future work.

6.2.3 Can the quality of user-generated tags for videos be evaluated and improved?

In Chapter 4 we demonstrate that game tags are not well suited for retrieving video fragments based on topic. This is mainly caused by the overwhelming presence of game tags which are not valid topical descriptors. In the second half of Chapter 4 we evaluate the quality of the game tags as topical descriptors with the intention to detect and filter out the ones that are not referring to topics. We explore several methods for predicting the tag quality — TF-IDF-based filtering, player reputation-based filtering, and topic models to name a few. The success of each method is judged against the evaluation dataset developed in the topical search study.

While the different methods that we studied performed with various success, the conclusion is that the game tags can be successfully exploited for topical video search provided there is a filtering process that would reduce or eliminate the effect of the non-topical tags. Our results show that after TF-IDF-based filtering the game tags can emulate the retrieval performance of the best performing system that utilizes manually crafted metadata for search. Also, combining TF-IDF filtered game tags with the manually crafted metadata yields an improvement of retrieval performance by 18%. The improvement is attributed to the increased retrieval recall stemming from the game tags. An important consequence of this result is that tagging games provide a cost-effective alternative for AV collection owners that do not possess the required manpower to manually annotate their material. This is of particular importance for the audio-visual cultural heritage domain. The national audio-visual archives throughout Europe have hundreds of thousands hours of digitized audio-visual material that is not readily accessible because of lack of annotations. The audio-visual archives can now launch Waisda?-like crowdsourcing campaigns confident that the collected tags will support tasks like visual and topical search. We come back to this point in Section 6.3.1.

6.3 Discussion and Further Research

In this thesis we investigate the added value of the tags collected with the video-labelling game Waisda? for video fragment retrieval. The general conclusion is that the game tags improve retrieval and on average perform
better compared to the manually-crafted metadata and closed captions. This statement comes with three caveats.

First, in our setting the taggers and the searchers originate from the same population which results in significant terminological overlap between the keywords used for search and the tags ascribed to the videos. We believe this to be a strong contributing factor for the success of the game tags for search in the context of our study. And while it may look like a limitation at first, in practice it is not. The terminology employed when tagging is general enough to cater to the needs of all but the expert searchers which may use more specialized (specific) terms when searching [30].

Second, in this work we considered only keyword-based search thus the said improvement in retrieval is limited to this modality. At the time of writing, other search modalities such as content-based video retrieval [120] are gaining more attention, however keyword-based search is still the most prominent one, especially on the Web.

Third, in this work we focussed on single domain (genre), non-fictional TV Comedy. Consequently, the obtained results are bound to this domain. However, the qualitative analysis of the tags revealed that they are usually general concepts describing visual objects and as such are not so tightly coupled with the genre of the video content they are describing. Our expectation is that the game tags will perform well for most of the non-fictional TV genres which appeal to a wider audience. It remains to be seen whether this will be the case for the fictional genres where more symbolism and abstract expressions can be expected. We hypothesise that without properly instructing the taggers or targeting a niche sub-community of experts the outcome will be less satisfactory.

6.3.1 A broader perspective

In the recent years there were advancements in many fields that are related to the scope of the work presented in this thesis. Following is a non-exhaustive list of the most relevant ones.

Crowdsourcing in Cultural Heritage Domain

In the recent years the bulk of the European cultural heritage has been undergoing a process of digitisation. A testament of that was the European Union funded project PrestoPRIME\(^2\) about digitisation and preservation of the European audio-visual heritage. The project’s consortium included

\(^2\)http://www.prestoprime.org/
all major European national audio-visual archives and among many of the
project’s deliverables was Waisda?. The project also delivered a digitisation
and preservation framework that enables the conversion of thousands of
hours of audio-visual material from analog to digital form. However, simply
by being in digital form and without descriptive metadata makes the audio-
visual collections only marginally more accessible than being on analog tapes.
Manual annotation by professional catalogers is prohibitively expensive. Ac-
tording to empirical evidence it takes usually around five times the duration
of a video for a cataloger to annotate it. In this thesis we have demon-
strated that GWAPs such as Waisda? provide a cost-effective alternative.
It is relatively inexpensive to deploy games such as Waisda? and after the
suitable processing the collected tags can be exploited for video retrieval.
This has important ramifications for the audio-visual archives. They can de-
ploy Waisda? and collect tags for the hundreds of thousands hours of video
material which are not annotated simply because it requires too much time
and effort for the catalogers to do it manually. The Sound and Vision in-
stitute, the Dutch national archive, already included the tags collected with
Waisda? in the in-house workflows and is cooperating with other institu-
tions in deploying Waisda? in various domains. In this thesis we developed
an extension module for Waisda? which enables seamless integration of the
game in any audio-visual collection workflow.

Waisda? is not the only initiative to introduce crowdsourcing in the cul-
ture heritage domain. Public Catalogue Foundation\(^3\) and BBC\(^4\) launched
Your Paintings.Tagger\(^5\) project to annotate the British national oil painting
collection (over 200,000 paintings) [29]. Your Paintings.Tagger is an online
crowdsourcing tool that enables the members of the public to tag the paint-
ings with the ultimate goal of making the collection more searchable. Another
crowdsourcing initiative is the OldWeather project\(^6\) where online volunteers
help scientists recover Arctic and worldwide weather observations made by
ships since the mid-19th century by transcribing ships’ logs. These tran-
scriptions will contribute to climate model projections and will improve our
knowledge of past environmental conditions [14]. Several other crowdsourcing
initiatives have been launched for transcribing old manuscripts [18, 75, 22]:

\(^3\)http://www.thepcf.org.uk/
\(^4\)http://www.bbc.com/
\(^5\)http://tagger.thepcf.org.uk/
\(^6\)http://www.oldweather.org
6.3. DISCUSSION AND FURTHER RESEARCH

Cymru1900Wales\(^7\), Transcribe Bentham\(^8\), Scripto\(^9\), DigitalKoot\(^{10}\), and Re-Script\(^{11}\) to name a few. Many other crowdsourcing initiatives, not necessarily in the cultural heritage domain, have been launched in the recent decade. We mention some of them in section 4.2. The driving force behind each of them, Waisda\(^?\) included, is to engage the online communities to create transformative knowledge. At the time of writing, Waisda\(^?\) is the only crowdsourcing project, that we are aware of, in the domain of audio-visual cultural heritage.

**Computer Vision**

Thanks to the advancements made in the area of deep learning \([9, 4, 96]\) the field of computer vision witnessed important breakthroughs in the recent years. Using vast number of high-resolution example images, the deep convolutional neural networks can be trained to classify and label scenes and objects depicted in new unseen images \([68, 31]\). An excellent example of this idea put into practice is the Microsoft’s Project Adam \([21]\). The Adam computer vision system is hailed to be advanced enough to tell the difference, for example, between a picture of a Pembroke and a Cardigan Corgi (dog breeds), a task that is quite difficult even for a person. The Adam system is fuelled by a deep convolutional neural network which has been trained with over 14 million images split up into 22,000 categories drawn from the ImageNet database\(^{12}\).

Another computer vision task that has benefited greatly from deep learning is face recognition. The Facebook’s DeepFace system is able to tell whether two previously unseen photos of faces show the same person with accuracy almost as high as that of a human regardless of variations in lighting or whether the person in the picture is directly facing the camera \([111, 138]\). The deep learning techniques have been applied to the medium of video as well. Karpathy et al. used convolutional neural networks for large-scale video classification. Deep learning has also been exploited to for human action recognition in videos \([101, 58]\).

In spite of all these advancements, the holy grail of computer vision, *total scene understanding*, is still well outside the reach or the current state-of-the-art systems. This is even more true for the medium of video which compared to still images has the additional temporal dimension. A direct

\(^7\)http://www.cymru1900wales.org/
\(^8\)http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham
\(^{10}\)http://www.digitalkoot.fi/
\(^{11}\)http://www.history.ac.uk/projects/digital/ReScript
\(^{12}\)http://www.image-net.org/
consequence of the temporal dimension and an excellent example of the complexity of scene understanding is the so-called Kuleshov effect. It is a cognitive phenomenon by which viewers derive more meaning from the interaction of two sequential shots than from a single shot in isolation [83]. This means that the meaning of a scene is not determined only by the scene itself but also by the scenes that preceded it. While the total scene understanding is presently unattainable for computers it is relatively easy for the average Waisda? player.

We have demonstrated that the Waisda? tags refer—among other things—to the more abstract aspects of the videos such as topics. This hints to a potential hybrid approach to video annotation that combines computer vision and a human-computation technique such as Waisda?. The computer vision system will detect and label objects, actions, generic scene classes, etc. and the Waisda? players will tag the more abstract aspect relating to the meaning of the scenes. In such a system there will be overlap, the computer vision system and the Waisda? players may ascribe the same tag. This redundancy is an affirmation of the correctness of the tag. A potential drawback of the computer vision systems in general is that they need to be trained to learn new concepts. Granted the Adam system can recognize over 22,000 concepts but this is still a finite and potentially limited set. The hybrid approach could be used to overcome this problem. The computer vision system can learn new concepts from the Waisda? tags. Indeed, the Waisda? tags and the associated scenes can be used as training examples for the computer vision system to learn the “unseen” concepts denoted by the tags. It remains to be seen whether the learned concepts by Adam system can be successfully used for retrieval.

Multimedia Information Retrieval

There are two major approaches to multimedia retrieval: text-based retrieval and content-based retrieval and according to Chu [23] the former approach is the most widely used in practice. Keyword-based search, which is a type of text-based retrieval, is the dominant search modality for images and videos on the Web. In this thesis we have demonstrated that the tags collected with Waisda? can be successfully exploited for keyword-based retrieval. This finding has important consequences for the accessibility of the videos on the Web where the size of the multimedia content has exploded exponentially in the recent years. According to the video marketing site ReelSEO\(^\text{13}\), as of December 2015, every minute 300 hours of video material is being uploaded on

\(^{13}\text{http://www.reelseo.com}\)
6.3. DISCUSSION AND FURTHER RESEARCH

YouTube\textsuperscript{14}[91] and YouTube is just one of the major video-sharing services. Waisda? and games alike can be used to annotate the vast number of videos on the Web and make them accessible via keyword-based search.

Although still an active research field, content-based retrieval has gained attention on the Web in the recent years. The major search engine Google\textsuperscript{15} introduced a reversed image search functionality called Google Images\textsuperscript{16}. Unlike traditional keyword-based image search, in Google Image the searcher submits an image as query and gets similar images and web pages as result. Bing\textsuperscript{17}, Yandex\textsuperscript{18}, and TinEye\textsuperscript{19} are other search engines which provide reversed image search that functions similarly as Google Images. There is plethora of other reverse image search systems: LAPIXA\textsuperscript{20}, picMatch\textsuperscript{21}, Pixsy\textsuperscript{22}, Mypicguard\textsuperscript{23}, ImageWiki\textsuperscript{24}, to name just a few.

We believe that content-based retrieval can benefit from video labelling games such as Waisda? and their output. One way is to combine the game tags with the content-based retrieval techniques into a hybrid approach. In fact, there is already research work done on the topic of combining high-level semantics such as text with content-based modalities [90, 136, 85, 55, 104]. Another way is to exploit the Waisda? tags to train supervised content-based retrieval methods. Waisda? tags are associated with a timepoint and deep-link to the image frame/video segment which they are describing. As such the tags and the associated image frames/video segments can serve as training examples.

6.3.2 Future Directions

Video labelling games are a human-based computation technique. Needless to say, anyone wishing to successfully deploy a Waisda? instance or other games alike must not neglect the human factor. Even the most perfectly executed video-labelling campaign from technological standpoint will fail if there are no taggers to contribute. In other words, marketing is vital. The empirical evidence collected from both Waisda? pilots supports this claim.

\textsuperscript{14}https://www.youtube.com
\textsuperscript{15}https://www.google.com
\textsuperscript{16}https://images.google.com/
\textsuperscript{17}https://www.bing.com/
\textsuperscript{18}https://www.yandex.com/
\textsuperscript{19}http://tineye.com/
\textsuperscript{20}https://lapixa.de/
\textsuperscript{21}http://www.picmatch.ca/
\textsuperscript{22}https://pixsy.com/
\textsuperscript{23}www.mypicguard.com
\textsuperscript{24}http://image.wiki/
On both occasions the bulk of the tags was accumulated during periods when there was an active marketing campaign for promoting the game. We believe that carefully applying the marketing mechanisms for acquiring new players by targeting the right online communities is a key factor for success. We mentioned already that annotation of content from certain genres such as art which may employ abstract symbolism will require more knowledgeable individuals. Our hypothesis is that engaging the right community leads to better tags. Understanding the relationship between target marketing and user engagement on one side and the quality of the resulting tags on the other is a worthwhile direction for future research.

What makes video annotation a difficult task is the fact that video is a medium that is extremely rich in meaning. The taggers can get overwhelmed by the complex interplay of objects and events especially in the fast-paced game setting. We believe that the game design directly influences the focus and quality of the resulting tags. For example, significant portion of the Waisda? tags refer to more noticeable aspects of the content such as moving objects in the background or in the foreground, or prominent stationary objects. This is hardly surprising, after all these are things that are easy to reach consensus on and ultimately that is the goal of the game from the perspective of the players. We hypothesise that if the underlying goals of the game creators are better aligned with the game objectives from the perspective of the players that will result in increased quality of the resulting tags. Understanding how the game design and game objectives influence the quality of the resulting tags is another worthwhile direction for future research.

We foresee that video-labelling games will be increasingly used for video annotation and eventually will become the dominant method due to their cost-effectiveness. Future research could focus on making them more effective by engaging the right communities and stimulating them to produce better tags by setting appropriate game objectives. That being said, the role of the catalogers and curators in future will be less about manual annotation and more about steering the annotation process — making strategic decisions about what kind of metadata to collect, what type of crowdsourcing campaign to launch, and which online communities to engage.

\footnote{An example goal could be tagging the occurrences of specific type of events.}
Appendices
Appendix A

Precision/Recall per Query for Visual Search

This appendix contains the detailed precision and recall metrics for each of the queries from the visual search study presented in Chap. 3. Note that the queries for which there were no results returned by the verified tags were left out.
APPENDIX A. PRECISION/RECALL PER QUERY FOR VISUAL SEARCH

- Precision/Recall at k for query: auto
- Precision/Recall at k for query: boer
- Precision/Recall at k for query: boerderij
- Precision/Recall at k for query: dochter
- Precision/Recall at k for query: dorp
- Precision/Recall at k for query: epe

Legend:
- ^D^ Recalls at k for all tags
- Recalls at k for verified tags
- Precision at k for all tags
- Precision at k for verified tags
APPENDIX A. PRECISION/RECALL PER QUERY FOR VISUAL SEARCH

Precision/Recall at k for query: gezond

Precision/Recall at k for query: huis

Precision/Recall at k for query: jan

Precision/Recall at k for query: kerk

Precision/Recall at k for query: kermis

Precision/Recall at k for query: kerst

Legend:
- Recall at k for all tags
- Recall at k for verified tags
- Precision at k for all tags
- Precision at k for verified tags
APPENDIX A. PRECISION/RECALL PER QUERY FOR VISUAL SEARCH

![Graphs showing precision/recall for different queries](image_url)
APPENDIX A. PRECISION/RECALL PER QUERY FOR VISUAL SEARCH

![Graphs showing precision and recall for different queries.](image-url)
Appendix B

Query Precision over Time for Visual Search

This appendix shows how the precision metric for each query changed over time as more tags are acquired. The precision metric figures originate from the visual search study presented in Chap. 3.
APPENDIX B. QUERY PRECISION OVER TIME FOR VISUAL SEARCH

Search precision of queries over time

Precision

Time (in Weeks)

Query Precision Over Time Graph
Appendix C

Query Recall over Time for Visual Search

This appendix shows how the precision metric for each query changed over time as more tags are acquired. The precision metric figures originate from the visual search study presented in Chap. 3.
APPENDIX C. QUERY RECALL OVER TIME FOR VISUAL SEARCH

Search recall of queries over time

Recall vs Time (in Weeks)
Appendix D

Precision/Recall per Query for Topical Search

This appendix contains the detailed precision and recall metrics for each of the queries from the topical search study presented in Chap. 4. Note that the queries for which there were no results returned by the verified tags were left out.
APPENDIX D. PRECISION/RECALL PER QUERY FOR TOPICAL SEARCH

![Precision/Recall for various queries.](image-url)
Precision/Recall at k for query: eten

Precision/Recall at k for query: feest

Precision/Recall at k for query: foto

Precision/Recall at k for query: friesland

Precision/Recall at k for query: gezin

Precision/Recall at k for query: gezond

- Recall at k for all tags
- Recall at k for verified tags
- Precision at k for all tags
- Precision at k for verified tags
APPENDIX D. PRECISION/RECALL PER QUERY FOR TOPICAL SEARCH

Precision/Recall at $k$ for query: huis

Precision/Recall at $k$ for query: kerk

Precision/Recall at $k$ for query: kip

Precision/Recall at $k$ for query: koffie

Precision/Recall at $k$ for query: mandy

Precision/Recall at $k$ for query: muziek

Legend:
- Blue dotted line: Recall at $k$ for all tags
- Green solid line: Recall at $k$ for verified tags
- Brown star line: Precision at $k$ for all tags
- Red square line: Precision at $k$ for verified tags
APPENDIX D. PRECISION/RECALL PER QUERY FOR TOPICAL SEARCH
APPENDIX D. PRECISION/RECALL PER QUERY FOR TOPICAL SEARCH
Bibliography


[8] Marek Bell, Stuart Reeves, Barry Brown, Scott Sherwood, Donny MacMillan, John Ferguson, and Matthew Chalmers. Eyespy: Support-


[30] Liliana Melgar Estrada, Michiel Hildebrand, Victor de Boer, and Jacco van Ossenbruggen. Time-based tags for fiction movies: Comparing experts to novices using a video labeling game. *Journal of the Association for Information Science and Technology*, accepted for publication.


on Image and Video Retrieval, CIVR ’10, pages 18–25, New York, NY, USA, 2010. ACM.


[85] Ximena Olivares, Massimiliano Ciaramita, and Roelof van Zwol. Boosting image retrieval through aggregating search results based on visual


Summary

As audio-visual archives become accessible to the wider internet audience, successful retrieval of archive items is becoming an increasingly challenging task. This is mainly caused by the lack of adequate annotations. Users cannot find what they are looking for because annotations are either not present or too expert-centric. Video tagging games are an attempt to alleviate this problem. Engaging the internet community to tag the videos addresses the issue of scarcity of annotations and yields tags that are community-centric. The focus of this thesis is on the role of game tags in audio-visual archives and their added value for video retrieval.

In the typical audio-visual archive ecosystem the game tags will coexist with professionally curated annotations. To understand what game tags bring to the table as opposed to the catalogers’ annotations we study the differences between them along two dimensions. The first dimension is the terminology: we perform a quantitative study to estimate the terminological gap between taggers and the catalogers. The second dimension is the descriptive scope i.e. which aspects of the video content are described and at what level of abstraction and granularity. We carry out a qualitative study in which we analyse the descriptive scope of the game tags. Our analysis shows that there is a terminological gap between the players of Waisda? and the professionals. This makes the game tags a valuable asset as they are from the users and for the users and can improve the access to the videos in the case when the ones accessing are the users themselves. Furthermore, we find that game tags predominately describe instances (objects) in the video and rarely scenes. This result provides important intuition about the usefulness of the game tags for video retrieval which is the focus of the remainder of the thesis.

Throughout this thesis we address two prominent search scenarios that arise in practice. The first scenario is retrieving video fragments that feature visual appearances of objects. In fact, this is very much like the instance search task from TRECVID with the difference that query is formulated in text and not by visual example as in TRECVID. The second scenario we
Summary

consider is *topical search* i.e. retrieval of video fragments that are about a given topic.

In Chapter 3 we study how effective are the game tags in addressing the first scenario. To this end, we design Cranfield-style experiments which require collection of “documents” (in this case video fragments tagged by users via Waisda?), set of queries which we derived from real-life query logs, and relevance judgments indicating which “documents” in the collection should be returned for each query. The results from the evaluation show that game tags are extremely well-suited for retrieving fragments that visually depict a given object of interest. They outperform the other types of metadata (including professional annotations) in pairwise comparison. We also establish that there is a difference in the focus and the scope of the game tags compared to the professional annotations; game tags being more fine-grained and professional annotations being more coarse-grained referring to the entire video and describing the prevalent topics. This property of the game tags makes them particularly appealing for business use-cases such as discovering/selling stock footage.

In Chapter 4 we evaluate the performance of game tags for topical search using another set of Cranfield-style experiments. The results demonstrate that raw, unprocessed game tags are not well suited for retrieving video fragments based on topic. While the search recall is satisfactory, the search precision leaves much to be desired. This is mainly caused by the presence of game tags which are not valid topical annotations. The fast pace of Waisda? usually results in players contributing mainly non-topical tags. To combat this effect in the topical search scenario, we characterize the quality of user tags as topical annotations and identify several tag features (filters) that serve as indicators whether game tags are useful for retrieval. Our results show that after filtering, game tags can emulate the retrieval performance of a baseline system that utilizes manually crafted metadata for search. An important consequence of this finding is that tagging games can provide a cost-effective alternative in situations when manual annotation by professionals is too costly.

The general conclusion of this thesis is that the game tags improve retrieval and on average perform better than the manually-crafted metadata and closed captions. This statement comes with three caveats. First, in our setting the taggers and the searchers originate from the same population which results in significant terminological overlap between the keywords used for search and the tags ascribed to the videos. We believe this to be a strong contributing factor for the success of the game tags for search in the context of our study. While this may look like a limitation at first, in practice it is often not, because the terminology employed when
tagging is typically sufficient to cater to the needs of all but the expert searchers which may use more specialized (specific) terms when searching.

Second, in this work we considered only keyword-based search thus the said improvement in retrieval is limited to this modality. At the time of writing, other search modalities such as content-based video retrieval are gaining more attention, however keyword-based search is still the most prominent one, especially on the Web.

Third, in this work we focussed on single domain (genre), non-fictional TV Comedy. Consequently, the obtained results are bound to this domain. However, the qualitative analysis of the tags revealed that they are usually general concepts describing visual objects and as such are not so tightly coupled with the genre of the video content they are describing. Our expectation is that the game tags will perform well for most of the non-fictional TV genres which appeal to a wider audience. It remains to be seen whether this will be the case for the fictional genres where more symbolism and abstract expressions can be expected. We hypothesise that without properly instructing the taggers or targeting a niche sub-community of experts the outcome will be less satisfactory.

The fact that game tags can successfully support video retrieval is of particular importance for the audio-visual cultural heritage domain. The national audio-visual archives throughout Europe have hundreds of thousands hours of digitized audio-visual material that is not readily accessible because of lack of annotations. The audio-visual archives can now launch Waisda?-like crowdsourcing campaigns confident that the collected tags will enable access to their collection items.
Samenvatting

Gezien het feit dat audiovisuele archieven steeds toegankelijker worden voor het grotere internetpubliek, wordt het succesvol ophalen van archiefitems een steeds uitdagender taak. Dit wordt voornamelijk veroorzaakt door het ontbreken van adequate annotaties: korte beschrijvingen van een video fragment. Gebruikers kunnen niet vinden waar ze naar op zoek zijn, omdat annotaties niet aanwezig zijn of teveel gericht zijn op het gebruik door experts. Video tagging games zijn een poging om dit probleem te verlichten. Door de internetcommunity in te schakelen om de video’s te taggen wordt het probleem van de schaarste aan annotaties aangepakt en worden tags gegenereerd die op de gemeenschap zijn gericht. De focus van deze thesis ligt op de rol van gametags in audiovisuele archieven en hun toegevoegde waarde voor het ophalen van video’s.

In het typische audiovisuele archief ecosysteem zullen de gametags naast professioneel samengestelde annotaties bestaan. Om te begrijpen wat gametags kunnen bieden in tegenstelling tot de annotaties van de documentalisten, bestuderen we de verschillen op twee dimensies. De eerste dimensie is de terminologie: we voeren een kwantitatief onderzoek uit om de terminologische kloof tussen taggers en de documentalisten in te schatten. De tweede dimensie is de beschrijvende scope, d.w.z. welke aspecten van de videoinhoud worden beschreven en op welk niveau van abstractie en granulariteit. Een kwalitatief onderzoek is uitgevoerd waarin de beschrijvende omvang van de gametags geanalyseerd worden. Onze analyse toont aan dat er een terminologische kloof is tussen de spelers van Waisda? en de professionals. Dit maakt de gametags een waardevolle toevoeging omdat ze afkomstig zijn van de gebruikers en voor de gebruikers zijn. Gametags kunnen de toegang tot de video’s verbeteren in het geval dat de gebruikers van het zoeksysteem en de spelers van de game uit dezelfde gebruikersgroep komen. Verder vinden we dat gametags voornamelijk concrete objecten in de video beschrijven en zelden scènes. Dit resultaat biedt belangrijke intuïtie over het nut van de gametags voor het ophalen van video’s, wat de focus is van de rest van de scriptie.
In deze thesis behandelen we twee prominente zoekscenario’s die in de praktijk ontstaan. Het eerste scenario is het ophalen van videofragmenten met visuele verschijningen van objecten. In feite lijkt dit erg veel op de “instantie” zoektaak van TRECVID met het verschil dat de zoekopdracht is geformuleerd in tekst en niet door een visueel voorbeeld zoals in TRECVID. Het tweede scenario dat we beschouwen, is zoeken naar videofragmenten die over een bepaald onderwerp gaan.

In hoofdstuk 3 bestuderen we hoe effectief de gametags zijn bij het behandel van het eerste scenario. Hiervoor ontwerpen we Cranfield-stijl experimenten waarvoor het verzamelen van “documenten” vereist is (in dit geval videofragmenten die zijn getagd door gebruikers via Waisda?), een set zoekopdrachten die we hebben afgeleid van zoekopdrachten in de praktijk en relevantieoordelens die aangeven welke “documenten” in de verzameling moeten worden geretourneerd voor elke zoekopdracht. De resultaten van de evaluatie laten zien dat gametags uitermate geschikt zijn voor het ophalen van fragmenten die een bepaald object van belang visueel weergeven. Ze presteren beter dan de andere typen metadata (inclusief professionele annotaties) in paarsgewijze vergelijking. We stellen ook vast dat er een verschil is in de focus en de reikwijdte van de gametags in vergelijking met de professionele annotaties; gametags zijn meer verfijnd, professionele annotaties gaan over de hele video en beschrijven alleen de belangrijkste onderwerpen. Deze eigenschap van de gametags maakt ze bijzonder aantrekkelijk voor zakelijke gebruikstoepassingen zoals het vinden en verkopen van “stock footage”.

In hoofdstuk 4 evalueren we de prestaties van gametags voor actuele zoekopdrachten. De resultaten tonen aan dat onbewerkte, onverwerkte gametags niet goed geschikt zijn voor het ophalen van videofragmenten op basis van het onderwerp. Hoewel de zoekopdrachten een bevredigend aantal antwoorden opleveren, laat de zoekprecisie te wensen over. Dit wordt voornamelijk veroorzaakt door de aanwezigheid van gametags die niet-relevante onderwerpen suggeren. Het snelle tempo van Waisda? resulteert meestal in spelers die alleen taggen wat ze direct zien. Om het zoeken op onderwerp te verbeteren, karakteriseren we de kwaliteit van gebruikertags als onderwerpbeschrijving en identificeren we verschillende tag-functies (filters) die dienen als indicatoren of game-tags die nuttig zijn voor het ophalen. Onze resultaten tonen aan dat gametags na filtering de ophaalprestaties van een baseline-systeem kunnen nabootsen, waarbij handmatig metadata worden gebruikt voor zoeken. Een belangrijk gevolg van deze bevinding is dat tagging-spellen een kosteneffectief alternatief kunnen bieden in situaties waarin handmatige annotatie door professionals te duur is.

De algemene conclusie van deze thesis is dat de gametags het zoeken verbeteren en gemiddeld beter presteren in vergelijking met de handmatig
Samenvatting

vervaardigde metadata en gesloten bijschriften. Deze verklaring komt met drie voorbehoorden.

Ten eerste zijn de taggers en dezoekers in onze context afkomstig van dezelfde populatie, wat resulteert in een aanzienlijke terminologische overlapping tussen de zoekwoorden die worden gebruikt voor het zoeken en de tags die aan de video’s worden toegeschreven. Wij geloven dat dit een factor is die sterk bijdraagt aan het succes van de gametags voor zoeken in de context van ons onderzoek. Hoewel het in eerste instantie een beperking lijkt, is dat in de praktijk niet het geval. De terminologie die wordt gebruikt bij het taggen is algemeen genoeg om tegemoet te komen aan de behoeften van iedereen behalve de ervaren gebruikers die wellicht meer gespecialiseerde (specifieke) termen gebruiken bij het zoeken.

Ten tweede, in deze scriptie hebben we alleen naar keyword-gebaseerde zoekopdrachten gekeken, dus de genoemde verbetering in het zoeken is beperkt tot deze modaliteit. Op het moment van schrijven krijgen andere zoekmodaliteiten, zoals het zoeken op basis van voorbeeld videofragmenten, meer aandacht, maar zoeken op basis van zoekwoorden is nog steeds de meest prominente, vooral op het web.

Ten derde richtten we ons in dit werk op single domain (genre), non-fictie tv-komedie. Bijgevolg zijn de verkregen resultaten gebonden aan dit domein. Uit de kwalitatieve analyse van de tags bleek echter dat dit meestal algemene concepten zijn die visuele objecten beschrijven en daarom niet zo nauw verbonden zijn met het genre van de video-inhoud die ze beschrijven. Onze verwachting is dat de gametags goed zullen presteren voor de meeste niet-fictionele tv-genres die een breder publiek aanspreken. Het valt nog te bezien of dit het geval zal zijn voor de fictionele genres waar meer symboliek en abstracte uitdrukkingen te verwachten zijn. We veronderstellen dat zonder de taggers behoorlijk te instrueren of een niche subcommunity van deskundigen te benaderen, de uitkomst minder bevredigend zal zijn.

Het feit dat gametags succesvol het ophalen van video’s kunnen onderteunen, is van bijzonder belang voor het domein van audiovisueel cultureel erfgoed. De nationale audiovisuele archieven in heel Europa hebben honderdduizenden uren aan gedigitaliseerd audiovisueel materiaal dat niet gemakkelijk toegankelijk is vanwege een gebrek aan annotaties. De audiovisuele archieven kunnen nu Waisda?-achtige crowdsourcingcampagnes lanceren, ervan overtuigd zijn dat de verzamelde tags toegang zullen geven tot hun collectie items.
<table>
<thead>
<tr>
<th>Year</th>
<th>Code</th>
<th>Candidate</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>1</td>
<td>Nicola Barile (UU)</td>
<td>Studies in Learning Monotone Models from Data</td>
</tr>
<tr>
<td>2014</td>
<td>2</td>
<td>Fiona Tuliyano (RUN)</td>
<td>Combining System Dynamics with a Domain Modeling Method</td>
</tr>
<tr>
<td>2014</td>
<td>3</td>
<td>Sergio Raul Duarte Torres (UT)</td>
<td>Information Retrieval for Children: Search Behavior and Solutions</td>
</tr>
<tr>
<td>2014</td>
<td>4</td>
<td>Hanna Jochmann-Mannak (UT)</td>
<td>Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation</td>
</tr>
<tr>
<td>2014</td>
<td>5</td>
<td>Jurriaan van Reijsen (UU)</td>
<td>Knowledge Perspectives on Advancing Dynamic Capability</td>
</tr>
<tr>
<td>2014</td>
<td>6</td>
<td>Damian Tamburri (VU)</td>
<td>Supporting Networked Software Development</td>
</tr>
<tr>
<td>2014</td>
<td>7</td>
<td>Arya Adriansyah (TUE)</td>
<td>Aligning Observed and Modeled Behavior</td>
</tr>
<tr>
<td>2014</td>
<td>8</td>
<td>Samur Araujo (TUD)</td>
<td>Data Integration over Distributed and Heterogeneous Data Endpoints</td>
</tr>
<tr>
<td>2014</td>
<td>9</td>
<td>Philip Jackson (UvT)</td>
<td>Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language</td>
</tr>
<tr>
<td>2014</td>
<td>10</td>
<td>Ivan Salvador Razo Zapata (VU)</td>
<td>Service Value Networks</td>
</tr>
<tr>
<td>2014</td>
<td>11</td>
<td>Janneke van der Zwaan (TUD)</td>
<td>An Empathic Virtual Buddy for Social Support</td>
</tr>
<tr>
<td>2014</td>
<td>12</td>
<td>Willem van Willigen (VU)</td>
<td>Look Ma, No Hands: Aspects of Autonomous Vehicle Control</td>
</tr>
<tr>
<td>2014</td>
<td>13</td>
<td>Arlette van Wissen (VU)</td>
<td>Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains</td>
</tr>
<tr>
<td>2014</td>
<td>14</td>
<td>Yangyang Shi (TUD)</td>
<td>Language Models With Meta-information</td>
</tr>
<tr>
<td>2014</td>
<td>15</td>
<td>Natalya Mogles (VU)</td>
<td>Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare</td>
</tr>
<tr>
<td>2014</td>
<td>16</td>
<td>Krystyna Milian (VU)</td>
<td>Supporting trial recruitment and design by automatically interpreting eligibility criteria</td>
</tr>
<tr>
<td>2014</td>
<td>18</td>
<td>Mattijs Ghijsen (VU)</td>
<td>Methods and Models for the Design and Study of Dynamic Agent Organizations</td>
</tr>
<tr>
<td>2014</td>
<td>19</td>
<td>Vincius Ramos (TUE)</td>
<td>Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support</td>
</tr>
<tr>
<td>2014</td>
<td>20</td>
<td>Mena Habib (UT)</td>
<td>Named Entity Extraction and Disambiguation for Informal Text: The Missing Link</td>
</tr>
<tr>
<td>2014</td>
<td>21</td>
<td>Kassidy Clark (TUD)</td>
<td>Negotiation and Monitoring in Open Environments</td>
</tr>
<tr>
<td>2014</td>
<td>22</td>
<td>Marieke Peeters (UU)</td>
<td>Personalized Educational Games - Developing agent-supported scenario-based training</td>
</tr>
<tr>
<td>2014</td>
<td>23</td>
<td>Eleftherios Sidirougos (UvA/CWI)</td>
<td>Space Efficient Indexes for the Big Data Era</td>
</tr>
<tr>
<td>2014</td>
<td>24</td>
<td>Davide Cebolin (VU)</td>
<td>Trusting Semi-structured Web Data</td>
</tr>
<tr>
<td>2014</td>
<td>25</td>
<td>Martijn Lappenschaar (RUN)</td>
<td>New network models for the analysis of disease interaction</td>
</tr>
<tr>
<td>2014</td>
<td>26</td>
<td>Tim Baarslag (TUD)</td>
<td>What to Bid and When to Stop</td>
</tr>
<tr>
<td>2014</td>
<td>28</td>
<td>Anna Chmielowiec (VU)</td>
<td>Decentralized k-Clique Matching</td>
</tr>
<tr>
<td>2014</td>
<td>29</td>
<td>Jaap Kabbedijk (UU)</td>
<td>Variability in Multi-Tenant Enterprise Software</td>
</tr>
<tr>
<td>2014</td>
<td>30</td>
<td>Peter de Cock (UVT)</td>
<td>Anticipating Criminal Behaviour</td>
</tr>
<tr>
<td>2014</td>
<td>31</td>
<td>Leo van Moergestel (UU)</td>
<td>Agent Technology in Agile Multiparallel Manufacturing and Product Support</td>
</tr>
<tr>
<td>2014</td>
<td>32</td>
<td>Naser Ayat (UVA)</td>
<td>On Entity Resolution in Probabilistic Data</td>
</tr>
<tr>
<td>2014</td>
<td>33</td>
<td>Tesfa Tegegne (RUN)</td>
<td>Service Discovery in eHealth</td>
</tr>
<tr>
<td>2014</td>
<td>34</td>
<td>Christina Manteli (VU)</td>
<td>The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems</td>
</tr>
<tr>
<td>2014</td>
<td>36</td>
<td>Joos Buiks (TUE)</td>
<td>Flexible Evolutionary Algorithms for Mining Structured Process Models</td>
</tr>
<tr>
<td>2014</td>
<td>37</td>
<td>Maral Dadvar (UT)</td>
<td>Experts and Machines United Against Cyberbullying</td>
</tr>
<tr>
<td>Year</td>
<td>Title</td>
<td>Institution</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------</td>
<td></td>
</tr>
</tbody>
</table>
| 2014-38 | Danny Plass-Oude Bos (UT)  
Making brain-computer interfaces better: improving usability through post-processing. |           |
| 2014-39 | Jasmina Maric (UVT)  
Web Communities, Immigration, and Social Capital |           |
| 2014-40 | Walter Omona (RUN)  
A Framework for Knowledge Management Using ICT in Higher Education. |           |
| 2014-41 | Frederic Hogenboom (EUR)  
Automated Detection of Financial Events in News Text |           |
| 2014-42 | Caresten Eijckhof (CWI/TUD)  
Contextual Multidimensional Relevance Models |           |
| 2014-43 | Kevin Vlaanderen (UU)  
Supporting Process Improvement using Method Increments |           |
| 2014-44 | Paulien Meesters (UVT)  
Intelligent Blauw: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden. |           |
| 2014-45 | Birgit Schmitz (OUN)  
Mobile Games for Learning: A Pattern-Based Approach |           |
| 2015-1 | Niels Netten (UVA)  
Machine Learning for Relevance of Information in Crisis Response |           |
| 2015-2 | Faiza Bukhash (UTV)  
Smart auditing: Innovative Compliance Checking in Customs Controls |           |
| 2015-3 | Twan van Laarhoven (RUN)  
Machine learning for network data |           |
| 2015-4 | Howard Spoelstra (OUN)  
Collaborations in Open Learning Environments |           |
| 2015-5 | Christoph Bösch (UT)  
Cryptographically Enforced Search Pattern Hiding |           |
| 2015-6 | Farideh Heidari (TUD)  
Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes |           |
| 2015-7 | Maria-Hendrike Peetz (UVA)  
Time-Aware Online Reputation Analysis |           |
| 2015-8 | Jie Jiang (TUD)  
Organizational Compliance: An agent-based model for designing and evaluating organizational interactions |           |
| 2015-9 | Randy Klassen (UT)  
HCI Perspectives on Behavior Change Support Systems |           |
| 2015-10 | Henry Hermans (OUN)  
OpenU: design of an integrated system to support lifelong learning |           |
| 2015-11 | Yongming Luo (TUE) (UVA)  
Designing algorithms for big graph datasets: A study of computing bisimulation and joins |           |
| 2015-12 | Julie M. Birkholz (VU)  
Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks |           |
| 2015-13 | Giuseppe Procaccianti (VU)  
Energy-Efficient Software |           |
| 2015-14 | Bart van Straalen (UT)  
A cognitive approach to modeling bad news conversations. |           |
| 2015-15 | Klaas Andries de Graaf (VU)  
Ontology-based Software Architecture Documentation |           |
| 2015-16 | Changyun Wei (UT)  
Cognitive Coordination for Cooperative Multi-Robot Teamwork |           |
| 2015-17 | André van Cleef (UT)  
Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs |           |
| 2015-18 | Holger Pirk (CWI)  
Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories |           |
| 2015-19 | Bernardo Tabuena (OUN)  
Ubiquitous Technology for Lifelong Learners |           |
| 2015-20 | Lois Vanhée (UU)  
Using Culture and Values to Support Flexible Coordination Using Culture and Values to Support Flexible Coordination |           |
| 2015-21 | Siibren Fetter (OUN)  
Using Culture and Values to Support Flexible Coordination Using Peer-Support to Expand and Stabilize Online Learning |           |
| 2015-22 | Zhemin Zhu (UT)  
Co-occurrence Rate Networks - Towards separate training for undirected graphical models |           |
| 2015-23 | Luit Gazendam (VU)  
Using Culture and Values to Support Flexible Coordination Cataloguer Support in Cultural Heritage |           |
| 2015-24 | Richard Berendsen (UVA)  
Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation |           |
| 2015-25 | Steven Woudenberg (UU)  
Bayesian Tools for Early Disease Detection |           |
| 2015-26 | Alexander Hogenboom (EUR)  
Sentiment Analysis of Text Guided by Semantics and Structure |           |
| 2015-27 | Sándor Héman (CWI)  
Updating compressed column stores |           |
| 2015-28 | Janet Bagorogoza (TIU)  
Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO |           |
| 2015-29 | Hendrik Baier (UM)  
Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains |           |
| 2015-30 | Kiavash Bahreini (OU)  
Real-time Multimodal Emotion Recognition in E-Learning |           |
| 2015-31 | Yakup Koç (TUD)  
On the robustness of Power Grids |           |
| 2015-32 | Jerome Gard (UL)  
Corporate Venture Management in SMEs |           |
| 2015-33 | Frederik Schadd (UM)  
Ontology Mapping with Auxiliary Resources |           |
| 2015-34 | Victor de Graaff (UT)  
Geosocial Recommender Systems |           |
| 2016-1   | Syed Saiden Abbas (RUN) | Recognition of Shapes by Humans and Machines |
| 2016-2   | Michiel Meulendijk (UU) | Optimizing medication reviews through decision support: prescribing a better pill to swallow |
| 2016-3   | Maya Sappelli (RUN) | Knowledge Work in Context: User Centered Knowledge Worker Support |
| 2016-4   | Laurens Rietveld (VU) | Publishing and Consuming Linked Data |
| 2016-5   | Evgeny Sherkhonov (UVA) | Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers |
| 2016-6   | Michel Wilson (TUD) | Robust scheduling in an uncertain environment |
| 2016-7   | Jeroen de Man (VU) | Measuring and modeling negative emotions for virtual training |
| 2016-8   | Matje van de Camp (TiU) | A Link to the Past: Constructing Historical Social Networks from Unstructured Data |
| 2016-9   | Archana Nottamkandath (VU) | Trusting Crowdsourced Information on Cultural Artefacts |
| 2016-10  | George Karafotias (VU) | Parameter Control for Evolutionary Algorithms |
| 2016-11  | Anne Schuth (UVA) | Search Engines that Learn from Their Users |
| 2016-12  | Max Knobbout (UVA) | Logics for Modelling and Verifying Normative Multi-Agent Systems |
| 2016-14  | Ravi Khadka (UVA) | Revisiting Legacy Software System Modernization |
| 2016-15  | Steffen Michels (RUN) | Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments |
| 2016-16  | Guangliang Li (UVA) | Socially Intelligent Autonomous Agents that Learn from Human Reward |
| 2016-17  | Berend Weel (VU) | Towards Embodied Evolution of Robot Organisms |
| 2016-18  | Albert Meroño Peñuela (VU) | Refining Statistical Data on the Web |
| 2016-19  | Julia Efremonova (TUE) | Mining Social Structures from Genealogical Data |
| 2016-20  | Daan Odijk (VU) | Context & Semantics in News & Web Search |
| 2016-21  | Alejandro Moreno Celleri (UT) | From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground |
| 2016-22  | Grace Lewis (VU) | Software Architecture Strategies for Cyber-Foraging Systems |
| 2016-23  | Brend Wanders (UT) | Repurposing and Probabilistic Integration of Data: An Iterative and data model independent approach |
| 2016-24  | Y. Kiseleva (TUE) | Using Contextual Information to Understand Searching and Browsing Behavior |
| 2016-25  | Dilhan J. Thilakaratne (VU) | In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains |
| 2016-26  | Wen Li (TUD) | Understanding Geo-spatial Information on Social Media |
| 2016-27  | Mingxin Zhang (TUD) | Large-scale agent-based social simulation: A study on epidemic prediction and control |
| 2016-28  | Nicolas Hönig (TUD) | Understanding Geo-spatial Information on Social Media |
| 2016-29  | Rurud Matthiasj (UVT) | The Eyes Have IT |
| 2016-30  | Mohammad Khelghati (UT) | Deep web content monitoring |
| 2016-31  | Elco Vriezekolk (UVT) | Assessing Telecommunication Service Availability Risks for Crisis Organisations |
| 2016-32  | Peter Bloem (UVA) | Single Sample Statistics, exercises in learning from just one example |
| 2016-33  | Dennis Schunselaar (TUE) | Configurable Process Trees: Ellicitation, Analysis, and Enactment |
| 2016-34  | Zhaochun Ren (UVA) | Monitoring Social Media: Summarization, Classification and Recommendation |
| 2016-35  | Andrea Minuto (UT) | Aligning Law and Action - a conceptual and computational inquiry |
| 2016-36  | Giovanni Sileno (UVA) | Materials that matter - Smart Materials meet Art & Interaction Design |
| 2016-37  | Merijn Bruijnes | Believable Suspect Agents: Response and Interpersonal Style Selection for an Artificial Suspect |
| 2016-38  | Spyros Martzoukos (UVA) | Combinatorial and compositional aspects of bilingual aligned corpora |
| 2016-39  | Christian Detweiler (TUD) | Accounting for Values in Design |
| 2016-40  | Thomas King (TUD) | Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance |
| 2016-41  | Saskia Koldijk (RUN) | Context-Aware Support for Stress Self-Management: From Theory to Practice |
| 2016-42  | Thibault Sellam (UVA) | Automatic assistants for database exploration |
2016-44 Bram van Laar (UT)  
Experiencing Brain-Computer Interface Control  
2016-45 Jorge Gallego Perez (UT)  
Robots to Make you Happy  
2016-46 Christina Weber (UL)  
Real-time foresight - Preparedness for dynamic innovation networks  
2016-47 Tanja Buttlar (TUD)  
Collecting Lessons Learned  
2016-48 Gleb Polevoy (TUD)  
The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains  
2017-1 Jan-Jaap Oerlemans (UL)  
Investigating Cybercrime  
2017-2 Sjoerd Timmer (UU)  
Designing and Understanding Forensic Bayesian Networks using Argumentation  
2017-3 Daniël Harold Telgen (UU)  
Grid Manufacturing: A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines  
2017-4 Murnal Gawade (CWI)  
Multi-core parallelism in a column-store  
2017-5 Mahdieh Shadi (UVA)  
Collaboration Behavior; Enhancement in Co-development  
2017-6 Damir Vandin (EUR)  
Intelligent Information Systems for Web Product Search  
2017-7 Roel Bertens (UU)  
Insight in Information: from Abstract to Anomaly  
2017-8 Rob Konijn (VU)  
Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery  
2017-9 Dong Nguyen (UT)  
Text as Social and Cultural Data: A Computational Perspective on Variation in Text  
2017-10 Robby van Delden (UT)  
(Steering) Interactive Play Behavior  
2017-11 Florian Kunneman (RUN)  
Modelling patterns of time and emotion in Twitter #anticipation  
2017-12 Sander Leemans (TUE)  
Robust Process Mining with Guarantees  
2017-13 Gjis Huisman (UT)  
Social Touch Technology - Extending the reach of social touch through haptic technology  
2017-14 Shoshannah Tekofsky (UVT)  
You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior  
2017-15 Peter Berck (RUN)  
Memory-Based Text Correction  
2017-16 Aleksandr Chuklin (UVA)  
Understanding and Modeling Users of Modern Search Engines  
2017-17 Daniel Dimov (UL)  
Crowdsourced Online Dispute Resolution  
2017-18 Ridho Reinanda (UVA)  
Entity Associations for Search  
2017-19 Jeroen Vuurens (TUD)  
Proximity of Terms, Texts and Semantic Vectors in Information Retrieval  
2017-20 Mohammadbashir Sedighi (TUD)  
Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility  
2017-21 Jeroen Linssen (UT)  
Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)  
2017-22 Sara Magliacane (VU)  
Logics for causal inference under uncertainty  
2017-23 David Graauw (UVA)  
Entities of Interest — Discovery in Digital Traces  
2017-24 Chang Wang (TUD)  
Use of Affordances for Efficient Robot Learning  
2017-25 Veruska Zamborlini (UVA)  
Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search  
2017-26 Merel Jung (UT)  
Socially intelligent robots that understand and respond to human touch  
2017-27 Michiel Joose (UT)  
Investigating Positioning and Gaze Behaviors of Social Robots: People’s Preferences, Perceptions and Behaviors  
2017-28 John Klein (VU)  
Architecture Practices for Complex Contexts  
2017-29 Adel Ahlurai (UTV)  
From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT  
2017-30 Wilma Latuny (UVT)  
The Power of Facial Expressions  
2017-31 Ben Ruijl (UL)  
Advances in computational methods for QFT calculations  
2017-32 Thaer Samar (RUN)  
Access to and Retrieval of Content in Web Archives  
2017-33 Armin van Loggem (OU)  
Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity  
2017-34 Maren Scheffel (OEU)  
The Evaluation Framework for Learning Analytics  
2017-35 Martine de Vos (VU)  
Interpreting natural science spreadsheets  
2017-36 Yuanhao Guo (UL)  
Shape Analysis for Phenotype Characterisation from High-throughput Imaging  
2017-37 Alejandro Montes Garcia (TUE)  
WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy  
2017-38 Jan-Jaap Oerlemans (UL)  
Investigating Cybercrime  
2017-39 Abdullah Kayal (TUD)  
Normative Social Applications  
2017-40 Sara Ahmadi (RUN)  
Understanding and Modeling Users of Modern Search Engines
Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
2017-41 Altaf Hussain Abro (VUA)
Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
2017-42 Adnan Manzoor (VUA)
Minding a Healthy Lifestyle; An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
2017-43 Elena Sokolova (RUN)
Causal discovery from mixed and missing data with applications on ADHD datasets
2017-44 Maaike de Boer (RUN)
Semantic Mapping in Video Retrieval
2017-45 Garm Lucassen (UU)
Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
2017-46 Bas Testenink (UU)
Decentralized Runtime Norm Enforcement
2017-47 Jan Schneider (OU)
Sensor-based Learning Support
2017-48 Yie Yang (TUD)
Crowd Knowledge Creation Acceleration
2017-49 Angel Suarez (OU)
Collaborative inquiry-based learning
2017-1 Han van der Aa (VU)
Comparing and Aligning Process Representations
2017-2 Felix Mannhardt (TUE)
Multi-perspective Process Mining
2017-3 Steven Bosems (UT)
Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
2017-4 Jordan Janeiro (TUD)
Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
2017-5 Hugo Huurdeman (VUA)
Supporting the Complex Dynamics of the Information Seeking Process
2017-6 Dan Ionita (UT)
Model-Driven Information Security Risk Assessment of Socio-Technical Systems
2017-7 Jieting Luo (UU)
A formal account of opportunism in multi-agent systems
2017-8 Rick Smetser (RUN)
Advances in Model Learning for Software Systems
2017-9 Xu Xie (TUD)
Data Assimilation in Discrete Event Simulations
2017-10 Julienka Mollee (VUA)
Moving forward: supporting physical activity behavior change through intelligent technology
2017-11 Mahdi Sargolzaei (VUA)
Enabling Framework for Service-oriented Collaborative Networks
2017-12 Xixi Lu (TUE)
Using behavioral context in process mining
2017-13 Seyed Amin Tabatabaei (VUA)
Computing a Sustainable Future: Exploring the added value of computational models for increasing the use of renewable energy in the residential sector
2017-14 Bart Joosten (UVT)
Detecting Social Signals with Spatiotemporal Gabor Filters
2017-15 Naser Davarzani (UM)
Biomarker discovery in heart failure
2017-16 Jaebok Kim (UT)
Automatic recognition of engagement and emotion in a group of children
2017-17 Jianpeng Zhang (TUE)
On Graph Sample Clustering
2017-18 Henricette Nakad (UL)
De Notaris en Private Rechtspraak
2017-19 Minh Duc Pham (VUA)
Emergent relational schemas for RDF
2017-20 Manxia Liu (RUN)
Time and Bayesian Networks
2017-21 Aad Slootmaker (OU)
EMERGO: a generic platform for authoring and playing scenario-based serious games
2017-22 Eric Fernandes de Mello Araujo (VUA)
Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks