On the Role of User-generated Metadata in Audio Visual Collections

Riste Gligorov
VU University Amsterdam
The Netherlands
r.gligorov@vu.nl

Michiel Hildebrand
VU University Amsterdam
The Netherlands
m.hildebrand@vu.nl

Jacco van Ossenbruggen
VU/CWI Amsterdam
The Netherlands
j.r.van.ossenbruggen@vu.nl

Guus Schreiber
VU University Amsterdam
The Netherlands
guus.schreiber@vu.nl

Lora Aroyo
VU University Amsterdam
The Netherlands
l.m.aroyo@vu.nl

ABSTRACT

Recently, various crowdsourcing initiatives showed that targeted efforts of user communities result in massive amounts of tags. For example, the Netherlands Institute for Sound and Vision collected a large number of tags with the video labeling game *Waisda?*. To successfully utilize these tags, a better understanding of their characteristics is required. The goal of this paper is twofold: (i) to investigate the vocabulary that users employ when describing videos and compare it to the vocabularies used by professionals; and (ii) to establish which aspects of the video are typically described and what type of tags are used for this. We report on an analysis of the tags collected with *Waisda?*. With respect to the first goal, we compared the the tags with a typical domain thesaurus used by professionals, as well as with a more general vocabulary. With respect to the second goal, we compare the tags to the video subtitles to determine how many tags are derived from the audio signal. In addition, we perform a qualitative study in which a tag sample is interpreted in terms of an existing annotation classification framework. The results suggest that the tags complement the metadata provided by professional cataloguers, the tags describe both the audio and the visual aspects of the video, and the users primarily describe objects in the video using general descriptions.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.

Keywords

tagging, video, tag analysis, professionals vs. end-users, games with a purpose

1. INTRODUCTION

Crowdsourcing has gained attention as a method to collect large numbers of metadata descriptions for media objects [2, 10, 15]. Based on the idea coined by Luis von Ahn [18], 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 [11] to integrate the game into their workflow to complement professional cataloguing and content based retrieval techniques [5]. 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 the paper is structured as follows. Section 2 discusses related work. Section 3 presents the approach we take in tackling the goals we set forth. Section 4 describes the materials we used in our study. Section 5 reports on the various experiments we performed on the user tags. Finally, section 6 draws conclusions and points to some further directions for research.

2. RELATED WORK

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 [18]. The first example of a GWAP was the ESP game [17], 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 [16], 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 [11]. 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 [9]. 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 [7]. 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.

2.2 Evaluation of end-user tags

The steve.museum research [10] 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 [1]. 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 folksomies. For example, in [4, 14] 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 pinpointed. Gruber [3] 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.

2.3 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 [12, 13] which focuses on the conceptual descriptions. Jaimes and Chang [8] developed 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. [6] 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.

2.3.1 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:

[1] http://www.videotag.co.uk/

[2] http://www.gwap.com/gwap/gamesPreview/popvideo/

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 [8] 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.

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 Google[3] 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 a 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 2.3.1.

4. MATERIALS

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

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 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.

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 [19]. 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 hyperonomy, hyponomy, meronomy etc. — which may be used across part of speech. Although Cornetto contains 59 different kinds semantic relations, hyperonymy and hyponomy are by far the most frequent ones, accounting for almost 92% of all semantic relation instances.

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 for select a subset. 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

[3]http://www.google.com

[4]Throughout this text we use the terms *player* and *user* interchangeably

Episode	All tags	Verified	Category
Farmer seeks wife 1	25,965	5,837	*Amusement*
Farmer seeks wife 2	22,792	6,153	*Amusement*
Traceless	1,007	274	*Amusement, Informative*
Reporter	403	73	*Informative*
The Walk	257	45	*Religious*

Table 1: Sample of waisda? episodes used in the experiments.

from a popular Dutch reality show, *Farmer seeks Wife*[5], categorized as amusement. The averagely tagged video is an episode from the *Traceless*[6] series, classified as amusement and informative program. The two low-tagged videos are episodes from *The Walk*[7] and *Reporter*[8] series, categorized as religious and informative, respectively. Table 1 summarizes the most pertinent information about the episodes. Prior research [1] 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.

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 the bottom of the screen[9]. Each dialog excerpt is accompanied with timepoints — 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].

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.

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. 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 professional cataloguer[11]. In addition, we observe that the verified tags are more often valid Dutch words than the non-verified ones.

	All tags		Verified	
Total	46,792		12,963	
In GTAA	3,850	(8%)	1,825	(14%)
In Cornetto	10,939	(23%)	5,669	(44%)

Table 2: Overlap of *Waisda?* tags with GTAA thesaurus and Dutch linguistic database, Cornetto.

	Facet	Tags
GTAA	Subject	1199
	Location	613
	Genre	52
	Person	118
	Maker	4
	Name	673
	Types	**Tags**
Cornetto	Noun	7222
	Verb	2090
	Adjective	1693
	Adverb	171

Table 3: Waisda? tags distribution over GTAA facets and Cornetto synset types.

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 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 1 provides a detailed view of these tags, showing the overlap between the different sets. We observe that 35% of the verified tags are

[5]http://www.bzv.kro.nl/
[6]http://spoorloos.kro.nl/
[7]http://dewandeling.kro.nl/
[8]http://reporter.kro.nl/
[9]Timed Text Working Group, http://www.w3.org/AudioVideo/TT/
[10]http://en.wikipedia.org/wiki/SubRip#SubRip_text_file_format
[11]GTAA contains all terms used to annotate videos in Sound and Vision

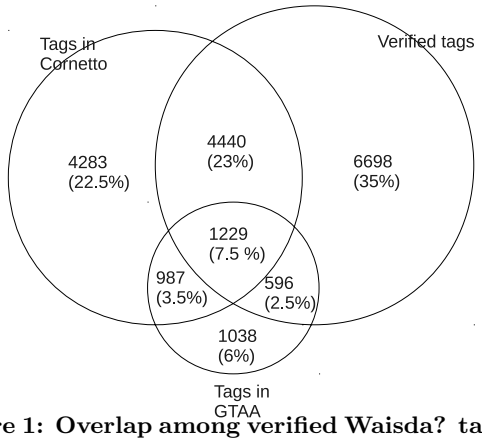

Figure 1: Overlap among verified Waisda? tags, tags in Cornetto, and tags in GTAA.

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 the vocabularies, but are not verified.

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 that 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.

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

Table 4: Overlap between the Waisda? tags and the video subtitles.

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 4.3 against the respective video subtitles for hearing impaired persons (see section 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 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 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.

5.3 Tag classification

In this experiment we performed a manual qualitative analysis on the tags of the five videos described in section 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 major-

(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).

	Abstract	General	Specific
Who	kind lady	woman	Anna
What	typical present	present	Amsterdam sausage
Where	idyllic countryside	kitchen	the Netherlands
When	elimination day	morning	May 10th 2008

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

Figure 2: Classification of user tags.

ity 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.

In classifying the conceptual object tags we followed the guidelines compiled by Hollink et. al [6] — figure 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 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 5.3 shows the distribution of the object-level tags across the categories of the Panofsky-Shatford model. 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

	Abstract	General	Specific	
Who	10	166	177	31%
What	73	563	12	57%
Where	0	68	8	7%
When	4	31	6	5%
	7%	74%	9%	

Table 5: Distribution of the object-level tags across the categories of the Panofsky-Shatford model.

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. [6]. 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%).

6. DISCUSSION AND FUTURE WORK

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 non-well-formed. 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.

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.

Together with The Netherlands Institute for Sound and Vision we are preparing a second pilot project with *Waisda?*. The results of this study show several limitations of the current metadata, that we aim to address in this pilot. One limitation is the low number of specific type of tags in the *who* and *where* facets. We are exploring 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. We are exploring if 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.

Another characteristic of the current *Waisda?* tags is that many are also found in the subtitles. In case these subtitles are also available for retrieval 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 labeling games is that only the verified tags are associated to the content as metadata. Our study shows that this approach 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. We are currently exploring the consequences of these methods.

Finally, in future work we will experiment with 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 do not provide a topical description of a fragment. We expect that the current tags are, therefore, suited to find objects within a specific video, but are as of yet less useful to find specific fragments. In future work we will explore methods to also collect topical descriptions of video scenes, by extending the game and/or with post-processing of the tags after the game.

Acknowledgements

We like to thank Johan Oomen, Maarten Brinkerink and Lotte Belice Baltussen from the Netherlands Institute for Sound and Vision for initiating and guiding the *Waisda?* project. We also like to thank Q42 for the development of *Waisda?* and making the collected data available. This research was partially supported by the PrestoPRIME project, funded by the European Commission under ICT FP7 (Seventh Framework Programme, Contract No. 231161).

7. REFERENCES

[1] L. B. Baltussen. Waisda? video labeling game: Evaluation report, 2009. [Online; accessed 20-January-2010] http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/.

[2] S. Chan. Tagging and searching — serendipity and museum collection database. In *Museums and the Web 2007: Proceedings*, Toronto, Canada, March 2007.

[3] T. Gruber. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems*, 3(2):1–11, 2007.

[4] M. Guy and E. Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1), January 2006.

[5] L. Hollink, B. Huurnink, M. van Liempt, J. Oomen, A. de Jong, M. de Rijke, G. Schreiber, and A. Smeulders. A multidisciplinary approach to unlocking television broadcast archives. *Interdisciplinary Science Reviews*, 34(2):257–271, June 2009.

[6] L. Hollink, G. Schreiber, B. J. Wielinga, and M. Worring. Classification of user image descriptions. *Int. J. Hum.-Comput. Stud.*, 61(5):601–626, 2004.

[7] B. Huurnink, C. G. M. Snoek, M. de Rijke, and A. W. M. Smeulders. Today's and tomorrow's retrieval practice in the audiovisual archive. In *ACM International Conference on Image and Video Retrieval*, 2010.

[8] A. Jaimes, R. Jaimes, and S. fu Chang. A conceptual framework for indexing visual information at multiple levels. In *in proceedings of SPIE Internet Imaging 2000*, pages 2–15, 2000.

[9] C. Jorgensen. Image access, the semantic gap, and social tagging as a paradigm shift. *Proceedings 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin*, 2007.

[10] T. Leason and steve.museum. Steve: The art museum social tagging project: A report on the tag con tributor experience. In *Museums and the Web 2009: Proceedings*, Toronto, Canada, March 2009.

[11] J. Oomen, L. Belice Baltussen, S. Limonard, A. van Ees, M. Brinkerink, L. Aroyo, J. Vervaart, K. Asaf, and R. Gligorov. Emerging practices in the cultural heritage domain - social tagging of audiovisual heritage. In *Proceedings of Web Science 2010: Extending the Frontiers of Society On-Line*. The Web Science Trust, April 2010.

[12] E. Panofsky. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Harper & Row, 1972.

[13] S. Shatford. Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, 6:39 – 62, 1986.

[14] L. F. Spiteri. Structure and form of folksonomy tags: The road to the public library catalogue. *Webology*, 4(2), June 2007.

[15] M. Springer, B. Dulabahn, P. Michel, B. Natanson, D. Reser, D. Woodward, and H. Zinkham. For the common good: The library of congress flickr pilot project: Report summary. Technical report, Library of Congress, 2008.

[16] R. van Zwol, L. Garcia, G. Ramirez, B. Sigurbjornsson, and M. Labad. Video tag game. In *WWW 2008*, April 2008.

[17] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM.

[18] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[19] P. Vossen, I. Maks, R. Segers, and H. van der Vliet. Integrating lexical units, synsets and ontology in the Cornetto database. In *LREC'08*, 2008.