Journal of the Association for Information Science and Technology

Time-based tags for fiction movies: Comparing experts to novices using a video labeling game

Journal:	*Journal of the Association for Information Science and Technology*
Manuscript ID:	JASIST-2013-08-0589.R4
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	27-Aug-2015
Complete List of Authors:	Melgar Estrada, Liliana; Carlos III University of Madrid, Library and Information Science Department Hildebrand, Michiel; Centrum Wiskunde & Informatica (CWI), de Boer, Victor; Vrije Universiteit Amsterdam, Computer Science van Ossenbruggen, Jacco; Vrije Universiteit Amsterdam, Computer Science; Centrum Wiskunde & Informatica (CWI),
Keywords:	subject experts < human information resources < (persons and informal groups), metadata < data < (documents by information content, purpose) < (document types), films (motion pictures) < video recordings < nonprint media < (documents by medium, physical form) < (document types)

Time-based tags for fiction movies:

Comparing experts to novices using a video labeling game

Liliana Melgar Estrada*

Carlos III University of Madrid, Library and Information Science Department, Spain. E-mail: lmelgar@bib.uc3m.es

Michiel Hildebrand

Centrum Wiskunde & Informatica (CWI), The Netherlands. E-mail: Michiel.Hildebrand@cwi.nl

Victor de Boer**, and **Jacco van Ossenbruggen**

Vrije Universiteit Amsterdam, The Netherlands. E-mail: {v.de.boer, j.r.van.ossenbruggen} @vu.nl

*Corresponding author

** Jacco van Ossenbruggen is also affiliated with Centrum Wiskunde & Informatica (CWI).

Abstract. The cultural heritage sector has embraced social tagging as a way to increase both access to online content and to engage users with their digital collections. In this paper, we build on two current lines of research. (1) We use *Waisda?,* an existing labeling game, to add time-based annotations to content. (2) In this context, we investigate the role of experts in human-based computation (*nichesourcing*). We report on a small-scale experiment in which we applied *Waisda?* to content from film archives. We study the differences in the type of time-based tags between experts and novices for film clips in a *crowdsourcing* setting. The findings show high similarity in the number and type of tags (mostly factual). In the less frequent tags, however, experts used more domain-specific terms. We conclude that competitive games are not suited to elicit real expert-level descriptions. We also confirm that providing guidelines, based on conceptual frameworks that are more suited to moving images in a time-based fashion, could result in increasing the quality of the tags, thus allowing for creating more tag-based innovative services for online audiovisual heritage.

Keywords: social tagging, tagging games, nichesourcing, expert tags, audiovisual heritage, film, *Waisda?*

Introduction

In the cultural heritage domain, social tagging has become an attractive solution to involve the public in the process of describing the objects in digital collections (Oomen & Aroyo, 2011). For example, the Steve museum social tagging project collected a large number of tags that describe artworks (Trant, 2009a). The *Waisda?* video labeling game, launched in 2009 by the Netherlands Institute for Sound and Vision, was used in two projects to collect tags for TV broadcasts and historic newsreels, showing that social tagging can also be applied to the audiovisual domain (Gligorov, Hildebrand, van Ossenbruggen, Schreiber, & Aroyo, 2011; Images for the Future, 2009). Together, the two projects resulted in over a million time-based tags that describe the content in the video, for example, depicted locations.

Analysis of the tags collected with *Waisda?* for TV broadcasts showed that users primarily describe the visual content at a general level (Gligorov et al., 2011). Motion pictures, however, have a distinctive form and a specific narrative (Bordwell & Thompson, 2003, p. 2) and involve different semantic dimensions compared to TV broadcasts, such as the use of framing, camera movements and composition to express meaning. Tags at this specific level are needed to describe adequately and retrieve film content, for instance, when users do archival footage research, based on "shot listings" (Turner, 2010; Wilkie, 1999). It is unclear if players of a video labeling game would provide specific tags of this kind.

In this paper we investigate the difference in the types of tags provided by experts and novices with three aims: 1) contributing to the understanding of the role of expert tags for content access in the audiovisual heritage domain, in line with the studies on *nichesourcing*; 2) continuing research on time-based metadata and labeling games initiated by the *Waisda?* experiments, exploring to what extent a video labeling game can be used to collect tags for films; and 3) contributing to the overall discussion of how social tagging can be applied to the film domain. By *film domain,* we mean mostly fiction movies, not necessarily celluloid films.

For this purpose, we designed a small-scale experiment using *Waisda?*, in which both film experts and novices performed time-based tagging for five film clips. This study does not seek generalizations, but identification of emergent issues in social tagging and human computation research applied to film images.

First, we present prior work related to our study. We describe the experimental design and setting and report our results and discuss them. We present the limitations of this study, followed by the main conclusions and ideas for future work.

Related work

We discuss four main topics related to our study: social tagging in the audiovisual heritage domain, tags from experts versus novices, guided tagging, and tag categories and models for image description.

Social tagging in the audiovisual heritage domain

Social tagging has been one of the earliest implemented collaborative practices for describing shared content online. Since in 2005 services like Furl, Flickr, and Del.icio.us started offering their users the option to add labels or tags to organize content (Smith, 2007), many websites have incorporated social tagging services, and research has not ceased in discovering new theoretical and practical approaches to this way of indexing digital information.

The cultural heritage sector has embraced this practice and is progressively incorporating it, together with other *crowdsourcing* initiatives, as part of their workflows (Oomen & Aroyo, 2011). However, regarding access to audiovisual heritage through socially generated tags, research is just starting.

State-of-the-art automatic moving image retrieval can achieve content-based indexing based on the images' low-level features and concept-based indexing based on derived high-level concepts (Stock, 2010). However, the performance is still not optimal to be used in all settings (Gibbon, Liu, Basso, & Shahraray, 2013; Yeh & Wu, 2014). In turn, different techniques for semi-automatic concept-based indexing at the shot level have been investigated by Turner (2009) though they only apply at a small scale. But socially generated tags (by niche groups and by the general crowd), if well guided, could help to bridge the gap 1) between content-based and concept-based annotations (as promulgated by Enser, 2000; and explored in Freiburg, Kamps, & Snoek, 2011; and Melenhorst, Grootveld, van Setten, & Veenstra, 2008) and 2) among concept-based annotations created manually (as different studies with tags have shown, such as Lu, Park, & Hu, 2010; Matusiak, 2006; and Springer et al., 2008).

In the audiovisual domain, social tagging research has focused mainly on recommendations of entire videos or movies based on tags and user profiles (for instance in the work by Bertini et al., 2013a, 2013b, and Gedikli & Jannach, 2013), and in video classification based on tags (for instance in Huang, Fu, & Chen, 2010). Little research exists, however, about the application of tags to time-based metadata, also called "time-coded metadata", or "strata" by Troncy, Huet, & Schenk (2011, p. 7), which is the information related to a specific time frame within video sequences. This research gap has been identified in Ballan, Bertini, Del Bimbo, Meoni, & Serra (2010; 2011); and Li et al. (2011), even though on a practical side, initial implementations of time-based social tagging are emerging in the audio domain, for instance the BBC's "Find, listen, label" tool for adding notes to radio programs (1).

The few exceptions to the lack of research in this area include an early study about tagging applied to the movie recommendation service "MovieLens" (Sen et al., 2006). Most related to our work are studies related to a larger effort to develop a framework for the *crowdsourcing* of film and television indexing by Geisler, Willard, & Whitworth (2010). Other related work consists of a study by Freiburg, Kamps, & Snoek (2011), that looks at the time-based metadata approach in combination with socially generated tags and automatically created annotations to video fragments of music concerts; the Larm Project in the radiophonic cultural heritage which gives prominence to user-driven annotations (Skov & Lykke, 2012), and the studies done in the framework of the *Waisda?* project.

Waisda? is a social tagging application and research project in the audiovisual heritage domain. Specifically it uses the idea of games-with-a-purpose (Ahn & Dabbish, 2008) to motivate users to contribute, since play and competition have been identified as motivating factors for tagging (Zollers, 2007). It was launched in 2009 by the Netherlands Institute for Sound and Vision. During the first pilot, the site received more than 12,000 visits, and attracted over 2,000 players, contributing 420,000 tags for 604 video items (Gligorov et al., 2011; Images for the Future, 2009). In the second pilot approximately 750,000 tags were collected. This is in line with the increasing popularity of human computation games (HCGs) for image description (Goh, Ang, Lee, & Chua, 2011; Goh & Lee, 2011). HCGs are one way of harnessing human intelligence, through the use of computer games, to perform activities that are impossible to automate, such as distinguishing types of fruits in an image (Goh et al., 2011). The first *Waisda?* pilots showed that *crowdsourcing*, in the form of a labeling game, can be also a good way to engage audiences with collections while obtaining content descriptors that can enhance retrieval (Gligorov, Hildebrand, van Ossenbruggen, Aroyo, & Schreiber, 2013).

In the film domain, content keywords have been utilized successfully for Fossati calls the "creative re-use of, or inspiration by archival material" (Fossati, 2009, p. 96). Examples of this approach are the "Celluloid Remix contest" (2), and "The Scene Machine" (3), which allow users to creatively explore online archival film footage relying upon keyword-based search, and to use existing labels to create their own content. However, these keywords are not socially generated but provided by the coordinating institutions. They also do not seem to be based on research about generating social tags in a moving image context.

However, there is consensus in that socially generated tags have quality problems associated with the use of non-words, polysemy, synonymy and lack of hierarchy (Guy & Tonkin, 2006; Matusiak, 2006; Lu et al., 2010), and to the lack of distinction of which type a tag corresponds to (Springer et al., 2008, p. 18). In the case of still image indexing, the existing problems for text indexing are even multiplied (Matusiak, 2006, p. 294) due to the semantic richness and ambiguity inherent to pictorial representations.

Nevertheless, it may be worthwhile to look for ways to surpass these disadvantages, since the application of social tagging may engage audiences and augment awareness of heritage collections (Springer et al., 2008), create different access points (Lu et al., 2010, p. 764; Thøgersen, 2013) that help increasing indexer-searcher consistency, and may complement automatic annotations (Freiburg et al., 2011). One initiative to improve tag quality is *nichesourcing*, a form of human computation which takes advantage of social tagging but involves experts, as opposed to *crowdsourcing*, in which taggers are the general public with no specific knowledge of a given domain (Boer et al., 2012).

Expert and novice generated tags

Social tagging has been defined as a way of organizing information by novices as opposed to the way indexing experts do (Peters, 2009, p. 1). One of the key factors in the success of social tagging in engaging different types of users is the reduction of intermediary steps followed in traditional indexing practices, saving the user from the need for first thinking on a concept and then representing it through the correct term from a controlled vocabulary (Halpin, Robu, & Shepherd, 2007). Different studies compare socially generated tags by non-expert users with the metadata created by indexing experts (Gligorov et al., 2011; Lu et al., 2010; Matusiak, 2006; Springer et al., 2008; Thøgersen, 2013; Trant, 2009b).

In our study, we focus on the relation between the types of generated tags and the participants' knowledge of the domain. Tsai, Hwang, & Tang (2011) looked at whether experts can provide a more consistent and representative set of tags for academic and scientific documents than novices, in the context of nanomaterial technology. They concluded that tags chosen by experts yielded better similarity and relevance values in all analyses and that these tags reflected better understanding of the content. Another study, in the radiological domain by Wang, Ni, Hua, & Chua (2012) explored how novices, intermediates and experts would describe medical images, finding that experts used more high-level image attributes that required high reasoning or diagnostic knowledge than novices, and that novices are more likely to describe basic objects that do not require much radiological knowledge. But Ådland & Lykke (2012) also found that tags can improve the interaction and communication between layman users and domain experts in a domain-specific setting (health information), helping to bridge between scientific terminology (and viewpoints) and everyday problems reflected in non-expert users' vocabulary.

Kang & Fu (2010) take this distinction a level further, by observing not only the tags or the tagging process of these two groups, but also the exploratory information search behavior of experts and novices using a social tagging system, in comparison to a general search engine. They found that expert-created tags could support the understanding of a topic by novices and increase their exploratory search. Closer to our research approach is a small-scale study by Darvish & Chin (2010), comparing film experts and novice tags in a video labeling setting, finding that expert tags were judged to be more relevant by both experts and non-experts and that non-expert viewers also created significantly better tags than the uploaders of the videos.

Guided tagging

For achieving consistency and quality in the tags, different studies explore mechanisms for guiding users in the tagging process. For instance, Smith (2007, p. 128) identified three categories of tag "suggestion systems": previously used tags (suggestions or recommendations based on a user's prior tags), popular tags (based on frequently used tags by others), and recommended tags, suggested by tagging systems based on their own criteria. Faceted tagging is another way of guiding the tagging process, by indicating different aspects of a resource that could be tagged (Smith, 2007, p. 76). For instance, Bar-Ilan, Shoham, Idan, Miller, & Shachak (2008, p. 941) found that structured tagging, which guides the user by presenting "fields" (such as "event, symbol, personality, date, place"), usually resulted in more detailed descriptions. In a practical application, the "Your Paintings" tagging project (4) applies this in practice: it guides users when tagging different aspects of a picture, such as things, people, places, events, subjects, and types. Sen et al. (2006) showed in an experiment on vocabulary formation in the Movie-Lens system how different design choices affect the nature/types of tags used, their distributions and the convergence within a group.

In sum, as Good, Tennis, & Wilkinson (2009, p. 14) point out, investigation on methods for guiding user contributions in particular directions is an important area of tagging behavior research. In the experiment we describe here, a group of randomly selected taggers received guidance in the form of instructional text informing about the types of tags which should be used.

Tag categories and models for image description

Although active research on tag categories exists (Peters, 2009, p. 196), to our knowledge, there are no studies about the different types of user-generated tags in a time-based fashion within the audiovisual domain. In our study, with the aim of creating an instructional guide on tag types for film content, and of observing semantic categories and types of tags used by expert and novice groups, we selected four types of tags by combining different models for fixed image analysis found in the literature. In its review, the Panofsky/Shatford matrix was found to be a widespread model for describing image content (Westman, 2009, p. 64). Panofsky (1977) addressed the levels of meaning in artistic images, defining three properties: pre-iconographical, iconographical and iconological. Layne (1986) followed with an extension of Panofsky's theory, adding four more facets (who, what, where, when).

Further, Hollink, Schreiber, Wielinga, & Worring (2004) adapted, extended and applied some of their preceding models for creating a framework that was used for classifying visual resources related queries and annotations. The framework distinguishes three viewpoints on images: the non-visual metadata level, the perceptual level, and the conceptual level.

More recently, Tirilly et al. (2012) proposed a model of image description based on characteristics obtained from experimental data in a study about the features of image similarity. According to them, their model provides a basis to define the image features that image retrieval systems should implement (p. 170). The features in their model refer to the image properties (e.g., type/technique, focus, point of view, lighting, contrast, file quality), to the scene's semantic and physical properties (e.g., place, time, color, composition), and to the objects' semantic and physical properties (e.g., nature, emotion, color, texture).

Golbeck, Koepfler, & Emmerling (2011) applied the Panofsky/Shatford model to the analyses of the social tagging behavior of image content. They tried to discover the relationship between tagging behavior and the features of the images which were tagged. They found that users' past experience with an image as well as the type of image being tagged creates significant differences in the number, order, and type of tags (p. 1750).

Even though the models mentioned above refer mainly to still-image analysis, they have been used to analyze moving images as well. Hollink (2006) used her framework for classifying visual resources (Hollink, 2006; Hollink et al., 2004) in three different contexts, one of them being broadcast news for a content-based image retrieval system. The results showed that the specific level was more important in the news domain than in the other domains (p. 121). In turn, Gligorov et al. (2011) used Hollink's and Panofsky/Shatford models in the analysis of *Waisda?* tags for television programs of a broad and entertaining nature.

In a study of key-frame extraction, Kim & Kim (2010) reviewed six representative models for still image analysis, concluding that people interact with images at three levels: primitive features (e.g., color and shape); derived attributes (e.g., specific objects), and semantic abstract attributes (e.g., the symbolic value) (Greisdorf & O'Connor, 2002)", which resemble the three panofskian levels.

In general, we found a lack of research about how these models for still image analysis can be applied or adapted to moving images, and observed a gap in the literature in identifying the formal and content attributes of time-based descriptions that are meaningful to expert and novice users.

Experimental design

Research questions

RQ1. How do film experts tag films compared to the general public? Do film experts, as opposed to novices, reflect their domain specific knowledge when tagging film content?

Tags are a spontaneous way to associate words with digital content, which reflect the users' personal understanding of a topic or their intentions with the digital resources (Tsai et al., 2011). For that reason, we might hypothesize that domain experts would use their domain-specific terminologies when tagging. We thus study the types of film experts' tags and compare the differences between film experts and novices when tagging film content in a realistic *crowdsourcing* environment. We analyze, among other things, the distribution of their respective contributed tags through different semantic levels.

RQ2. Can we influence the type of time-based tags that users enter with specific instructions?

One of the problematic issues of indexing/tagging audiovisual content is that there are many levels or dimensions of meaning involved. To address this question, we investigate if experts and novices enter more specific tags when they receive instructions of using different semantic categories that may apply to film content.

Test procedure

To address our research questions, we designed a 2 × 2 between-subject study for which two groups of participants were selected: film experts and domain novices. In turn, these groups were divided into two sub-groups: one having instructions (guidance in which types of tags they could use), and the other one having only general indications on how to play the game, but no instructions on the types of tags to enter. .

All participants were asked to play a game with each of the five videos. Since we were interested in the types of tags, they were allowed to use their mother tongue when tagging if it was English, Dutch or Spanish, with the aim of favoring their spontaneity. The participants were asked to fill in a questionnaire after completing the five games.

Selection of participants

In total 36 persons participated in this study: 18 film experts and 18 domain novices, 9 out of the 18 in each group received instructions and 9 did not. The participants were selected in two different ways:

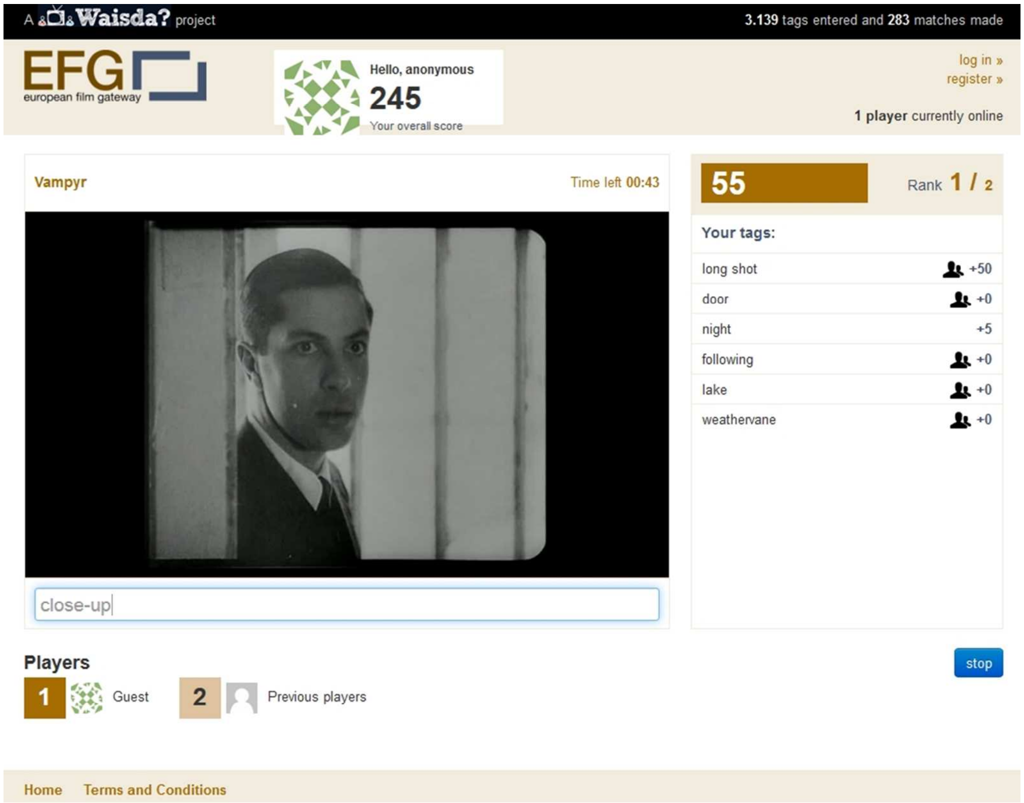
- **The film experts**. We considered people involved with film content at a professional or academic level and linked to film-related institutions. Our participants were contacted in film and television archives, universities, a government institution, and at a national library's film archive. They were based in The Netherlands, Norway, United States, Spain, and Colombia. In total, 45 invitations were sent, and 18 experts completed the full experiment (response rate: 40%). This group included participants who were film historians (scholars), cataloguers or archivists (curators), filmmakers, film/video technicians and film programming staff. All of them had an academic background in and/or formal education related to cinema. The age of the experts was between 30 and 39 (n=12), 50 and 59 (n=3), 20 and 29 (n=2), and 40 and 49 (n=1). Half of the participants had working experience with film materials and content of 10 years or more (n=9); between 7 and 9 years (n=6), 4 to 7 years (n=2), and one was a junior researcher (less than 3 years of working/research experience). There were twelve females and six males.
- **Film novices.** As non-domain experts, we considered people without a professional or academic relation to film content, and people not familiar with terminologies related to film. They were recruited by using an informal call for participation on one of the author's Facebook pages, indicating that not being a film expert or enthusiast was the only requirement. In total, we got 26 positive replies. From those, 18

completed the full experiment.

The novice group consisted of professionals with high-level education, mainly with a Library and Information Science background. This indexing expertise factor was not intentionally sought in the study, but since we were interested in domain specific knowledge we did not consider it a problem, rather we saw it as an advantage, since it helped us have a higher number of participants in all groups with knowledge and experience with tags and keywords. Regarding their ages, most novices were between 30 and 39 (n=9), the others were between 20 and 29 (n=5), 40 and 49 (n=2), and 50 and 59 (n=2). All novices defined themselves as such, that is, their domain-specific knowledge or distinct concern about films was null, and their interest in them was not explicitly reported to go beyond occasional movie-going activities. There were fourteen females and four males.

Prototype application

We used the *Waisda?* system (5) for the experiment setup. This is available as free and open source software at the GitHub repository (6). Figure 1 shows a screenshot of the tagging interface where it is possible to see how tags are entered while the video plays, being attached to a specific time point in the video. Users get points by entering tags, and a higher score when the tags match with the tags entered by other participants. A detailed explanation of the software, game rules and interface is described by Hildebrand et al. (2013).



long shot	+50
door	+0
night	+5
following	+0
lake	+0
weathervane	+0

close-up

Figure 1. Waisda-Efg tagging interface snapshot

For the experiment, the functionality of *Waisda?* was modified in two ways. Firstly, we neutralized the effect of the game scores on the tagging behavior: points were not given when the tags entered by one participant matched with tags entered by other participants. This was done to prevent participants from entering the types of tags that will maximize their score. For example, if a player observes that by entering 'woman', (s)he is rewarded with points, then (s)he would be encouraged to enter other tags of that type, such as 'man', 'dog', etc. This is what Fu, Kannampallil, Kang, & He (2010) called "semantic imitation", where "users who can see tags created by others tend to create tags that are semantically similar to these existing tags". Semantic imitation is an important characteristic of tagging games, but for the purpose of our experiment it had to be neutralized. As a solution, we decided to retain the scoring mechanism of the game, but to control the tags that are rewarded with points, in order to guarantee a fair distribution over the different tag category types (see "user instructions" section). Therefore, we introduced a single non-real player (a bot) that all the participants competed against, exerting the same influence on all subjects. The players were rewarded with points for matching with tags of the bot, but were unaware that they were not competing with other players. For each of the five film clips, we created a set of tags for the bot that covered each of the five tag categories included in the instructions. In this way, participants were rewarded for entering matching tags in the different categories, and not only for factual tags, which was the most common tag type in previous *Waisda?* experiments.

As a second modification, we disabled the display of tags entered by other players in the *Waisda?* game and on the *Waisda?* homepage, to neutralize all tag suggestions other than the instructions of the experiment.

Selection of film clips

We uploaded five clips from the European Film Gateway (EFG) (7) into the system. The EFG is a portal that provides access to the digitized collections of around sixteen European film archives and cinémathèques. We made a purposive sampling by selecting five clips according to the following criteria:

- They should be from films with no dialogs, with the aim of avoiding script transcription as much as possible;
- They should be short (no longer than five minutes), as previous *Waisda?* studies had indicated that the players prefer playing games with short clips.

Except for a Swiss short film, our final selection included movies from renowned Danish and German film classics or directors; we also assumed that if these movies were presented at the EFG their value was previously assessed. The five selected film clips were (clip duration is between brackets, and a link to the EFG record is referenced): “Den flyvende cirkus” (Alfred Lind, Denmark, 1912; [02:02]) (8), “Die Gezeichneten” (Carl Th. Dreyer, Germany, 1922; [00:37]) (9), “L’aiguille” (William Piasio, Switzerland, 1961, [05:55]) (10), “Metropolis” (Fritz Lang, Germany, 1926, [01:30]) (11), and “Vampyr” (Carl Th. Dreyer, Germany/France, 1932, [01:36]) (12).

Participants’ instructions

All participants received a common set of instructions by email, indicating how to play *Waisda?* (also available on the Waisda/EFG homepage). Participants that were part of the “instruction group” received another set of instructions, with details on the types of tags they could use (see “classification No.1” in the “Data analysis procedures” section). We created a simple “instructional model” based on some features of the models described in the section “tag categories and models for image description”. The following were the resulting instructions that we provided to the participants:

“Tags consisting of one or two words are more likely to match than longer phrases. Tags may be about the following aspects (please try to cover as many as you can during the game):

- **Facts**. What you see or hear in the scene, such as objects, persons, places and actions (e.g. woman, sofa, London, R2D2, murder).
- **Cinematography**. Stylistic features, such as form, style, framing, camera movement, lightning key, type of shot, camera angle (e.g. backlighting, wide-angle, caligarism).
- **Explanations**. Symbolic interpretation of the meaning or theme (e.g. psychotic rage, oppression, dehumanization).
- **Emotions**. The emotions, thoughts or intentions of the characters (e.g. bored, despair) or your own emotions (e.g., fascinating).
- **Other.** You can use other types of tags that are not described here".

We did not intend to create a "new" model or set of categories in this text, but rather interpreted and summarized some of the important features pointed in the existing models for image analysis related to film content. For instance, the "Facts" category, is inspired by Panofsky-Shatford's 'pre-iconography/ generic 'of' and Iconography / specific 'of'', and in Baca's (2002) 'Ofness' categories. Our "Emotions" concept coincides with Panofsky's (1955) 'Pre-iconographic (expressional) category' and other models which consider emotional abstraction (Eakins, Briggs, & Burford, 2004). Our "Explanation" type was derived from Panofsky's (1977) 'Iconology' category and Ingwersen's (1992) 'aboutness', and our "Cinematography" type from Hollink, Schreiber, Wielinga, & Worring's (2004) 'perceptual' category and from one of the key books on cinematography (Bordwell & Thompson, 2003).

Questionnaire

The participants were asked to fill in a questionnaire after completion of the test (13). The questionnaire consisted of 22 questions, divided into three sections: demographic information and expertise level; previous experience with indexing, tagging and labeling games; and the participant's experience with the game and experiment. In this last set of questions, participants were asked to rate their level of difficulty in coming up with tags, the influence that scoring in the game had on their motivation, the usefulness of the instructions, and their perception of the value of their tags for future use. The participants were also asked to select the types of tags (factual, emotional, etc.) they used, according to their judgment. There were also open questions in which participants could write their comments about these different aspects.

Data analysis procedures

We omitted tag stemming procedures since we are mainly interested in the type of tags that were entered, and not in the matching tags or tags morphology. All tags entered in Spanish and Dutch were manually translated into English, and misspellings were corrected, only with the aim of facilitating the tag category analysis (14).

In the quantitative analysis of the tags, we consider the number of tags that were entered. In this experiment, we do not include precise quantitative results of matching tags, due to the presence of tags in different languages. In the semantic analysis of the tags, in order to analyze their types, we manually classified them according to four different tag classifications (Classification No.1 corresponds to the instructions given to the participants, while Classifications No.2 to 4 were used for complementing the analysis but were not provided to the participants. In these last three classifications, we followed the same approach as in Gligorov et al. (2011)):

- *Classification No.1***.** “Instructional model” (**Facts**, **Emotions**, **Explanations**, **Cinematography**, **Other**). For the criteria to classify a tag in these categories, we used the examples and descriptions given to the participants, and we added some criteria for classifying the data.

- *Classification No.2*. “Hollink’s model”, also as used in Gligorov et al. (2011) includes the **Non-visual level** (descriptions that are meant to describe the context of the video but not its content); the **Perceptual level** (tags that are derived from low-level audio and visual features of the video); and the **Conceptual level** (tags that describe the content of the image, giving information about the semantic content of the image). We only use this classification to filter out the conceptual tags.

- *Classification No.3***.** “Panofsky-Shatford model (specific, abstract, general),” as used in Gligorov et al. (2011). At this level, tags that were classified as conceptual are classified according to their specificity level into specific, abstract or general. **Specific** (iconography) tags possess the property of uniqueness, for example, the name of a person or place. **Abstract** (iconology) tags are those which level of subjectivity allows for differences in opinion, (e.g., ‘crazy woman’). Also tags expressing relationships (e.g., ‘friends’), or tags related to occupations (e.g., ‘artist’). The last category is **General** (pre-iconography), which can be derived from the visual properties of the image or sequence alone. Tags classified as General do not have to be correct (‘dog barking’, and ‘duck quacking’ were used in the same time frame, this low level of subjectivity is not enough to consider the tag Abstract).

- *Classification No.4***.** "Panofsky-Shatford model (who, what, where, when)." We used the concepts from Shatford (1986): "**Who**" refers to the concrete objects and beings, animated or inanimate; or individually named persons, animals, things; or to kinds of persons, animals, things; or to mythical beings, abstractions manifested or symbolized by objects or beings. "**Where**" to a location and "**When**" to time. "**What**" refers to an event in the video: "what are the objects and beings doing? (action, events, emotions)", explains Shatford.

The tags were manually classified by one of the authors. A sample of the tags was classified by a second person for assuring the consistency of the classification criteria. We used a quota sample by randomly selecting tags created by each of the four subgroups for each video. The Cohen's kappa (k)[2] was used as a measure of agreement between both annotators. The results were reasonable for three of the classifications (0.67 for 1 and 2, and 0.62 for classification 3). The agreement was low (0.32) for classification 4. However, more in-depth analysis showed that this was due to a different interpretation of the Panofsky-Shatford's model in relation to the "Who" and "What" categories, which are explained differently in the original Shatford (1986) model, and in Gligorov et al (2011, p. 150). This does not reflect a disagreement in the tags classification but a different interpretation of the model. Since it was applied systematically in the classification by each annotator of a small proportion of tags, we concluded that the categorization was consistent and not arbitrary and that we could use it for analyzing our results.

After tag classification procedures, we manually clustered synonyms and singular/plural forms to look at the most frequent types of tags from a semantic perspective (the tags obtained from these clusters were used in Table 2 and 4). Finally, we analyzed the answers to the questionnaire: 1) to help interpret the results of the quantitative and semantic analysis, and 2) to discover the participants' perceptions of tagging behavior.

Findings and discussion

Next, we present the findings from the examination of the tags and the analysis of the questionnaire answers.

Number of tags

The 36 participants contributed a total of 2,943 distinct tag entries for the five videos. 2,404 were in English, 262 in Spanish, and 276 in Dutch. From the 2,404 English tags, 1,137 were unique. **Table 1** shows the means and standard deviation of the tags entered by each group. The high standard deviation among the participants in group D (58.1) was due to the presence of one “super-tagger” (as called by Trant (2009b). However, we did not detect any outliers (using the outlier labeling rule with a value of 2.2 as the multiplier).

Group	N	Total tags	Mean	Median	Min	Max	Standard deviation
A. Experts/ No instructions	9	641	71.2	66.0	27	140	40.9
B. Experts/ instructions	9	773	85.89	77.0	48	140	28.17
C. Novices/ No instructions	9	738	82.0	61.0	23	193	58.1
D. Novices / instructions	9	791	87.9	88.0	55	150	31.0

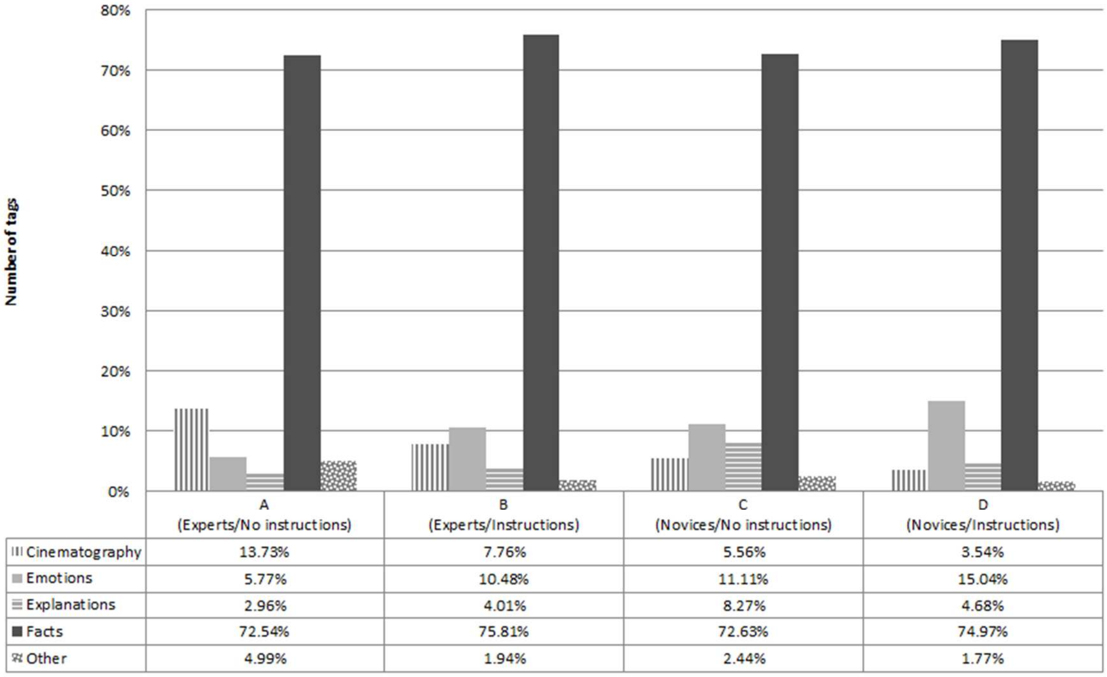
Table 1. Descriptive statistics of the number of tags per group (5 film clips, total duration: 700 sec.).

A Kolmogorov-Smirnov test showed that tags per group and video were not normally distributed. We therefore chose to conduct a Kruskal-Wallis test (a non-parametric test for independent samples and three or more groups) to examine the relationship between number of tags, expertise and instructions among all groups, as well as a Mann–Whitney U test for testing differences between pairs of groups.

The results showed that, in most cases, there is no effect of expertise and/or instructions in the number of tags entered by the different groups ($p > 0.05$). One exception appears in the evaluation at the individual video level, for which there was a significant difference for the clip of “Metropolis”: i) in the number of tags entered between all groups ($p = 0.013$); ii) between the groups A and C ($p = 0.019$); and iii) between the groups B and D ($p = 0.024$). We will comment on this later.

Types of tags

To observe the types of tags among the different groups, we used ”Classification No.1”. As we can see in **Figure 2**, the distribution of the types of tags among the different groups shows that all of them predominantly entered “Factual” tags. To illustrate which tags belong to each category, **Table 2** includes the three most frequent tags per group**.**



	A (Experts/No instructions)	B (Experts/Instructions)	C (Novices/No instructions)	D (Novices/Instructions)
▥ Cinematography	13.73%	7.76%	5.56%	3.54%
▒ Emotions	5.77%	10.48%	11.11%	15.04%
▤ Explanations	2.96%	4.01%	8.27%	4.68%
■ Facts	72.54%	75.81%	72.63%	74.97%
▩ Other	4.99%	1.94%	2.44%	1.77%

Figure 2. Proportional distribution of tags types across different categories (Classification No.1), percentage in relation to the total tags per group.

Categories	A (Experts/No instructions)	B (Experts/Instructions)	C (Novices/No instructions)	D (Novices/Instructions)
Cinematography	silent film; black and white; fiction	silent film; black and white; close-up	black and white; silent film; drama	black and white; silent film; close-up
Emotions	mystery; danger; fear	danger; help; angry	old; pain; scary	fear; relief; anger
Explanations	rebellion; expressionism; dystopia	expressionism; death; poverty	death; impressionism; luck	lucky; death; menacing music
Facts	door; train; smoking	shadow; smoking; monkey	shadow; workers; train	shadow; monkey; bell
Other	film; dreyer; german	german; vampyr; early cinema	german; vampyr; italy	german; metropolis; french

Table 2. Three most frequent tags in each category of Classification No.1 per group.

"Factual" tags correspond to objects or actions that are depicted in the scenes. These "ofness" words (as defined by Baca (2002); Peters (2009); Layne (1986)) correspond to what Panofsky calls the "pre-iconographical" level of meaning: the description of "primary or natural subject matter", which is apprehended by identifying *pure forms* (Panofsky, 1977, p. 5). Even though object identification is not a simple process (from the semiotic point of view), it is assumed here that these descriptions do not require film domain specific knowledge.

To examine closer what happened in the other four tag categories, and for observing the effect of expertise and instructions in the distribution of the types of tags, we performed a Kruskal-Wallis test again, and a Mann–Whitney *U* test for testing differences between pairs of groups. **Table 3** shows the cases with a statistically significant difference (*p* < 0.05).

	All groups (A, B, C, D)	Experts (No instructions/ Instructions) (A and B)	Novices (No instructions/ Instructions) (C and D)	Experts and Novices (No Instructions) (A and C)	Experts and Novices (Instructions) (B and D)
CINEMATOGRAPHY	0.102	0.340	0.161	0.387	0.024
EMOTIONS	0.001	0.024	0.031	0.003	0.113
EXPLANATIONS	0.338	0.931	0.050	0.136	0.666
FACTS	0.498	1.000	0.190	0.605	0.666
OTHER	0.383	0.222	0.387	0.436	0.546

Table 3. p values from Kruskal-Wallis and Mann–Whitney U test considering the five film clips. Cells in grey scale indicate a statistically significant difference at the p<0.05 level.

In **Table 3** we observe that there is a significant difference in the use of tags of the type "Emotion" between all groups, and by almost all the analyzed pairs of groups. This result was not expected. The group of experts with no instructions (A) had significantly fewer tags of the type **"Emotions"** than the respective novices group (C) (5.77% vs. 11%, *p*=0.003). In turn, the groups with instructions (B and D) entered more tags of this type than their counterpart with no instructions (A and C) (10.48% vs 5.77% *p*=0.024 for the experts groups, and 15% vs 11%, *p*=0.031 for the novices groups).

This difference may be caused by the level of awareness that the instructed groups gained on this type of tag. "Emotional" tags correspond to feelings expressed by the characters in the scenes as detected by the taggers (e.g. 'angry'), or to feelings experienced by the tagger her/himself (e.g. 'creepy'). The last type coincides with what Zollers (2007) identified as "opinion tags". Normally, the use of emotional attributes is not prescribed by traditional cataloguing or indexing guidelines. However, there is growing interest in the structured identification of emotional aspects from various art forms for different purposes (e.g., movie recommendation). Affective tagging could be used as part of user engagement activities (e.g., as in the "Emolab project" (15)), and/or for retrieval based on non-factual information during footage finding or research. For instance, Inskip, MacFarlane, & Rafferty (2008) describe the process of searching for accompanying music to film scenes, which involves highly subjective affective meanings, where motional tags could be useful. In turn, there is active research in the psychology domain (Bálint & Kovács, 2012) and in film studies (16) about the emotional involvement of the film viewer, which require or benefit from this type of tagging.

Also, in **Table 3** we can observe a predictable result in relation to "Cinematographic" tags: a significant difference (*p*= 0.024) in the number of tags entered by group B in relation to group D (7.76% vs. 3.54% of each group's total tags, as it can be seen from the proportions in **Figure 2**). **"Cinematographic"** tags correspond to domain-specific terms, such as photographic aspects of the shots or framing, camera movements or editing characteristics. In relation to RQ1, on whether experts' tags reflect their specific knowledge, we expected that the lack of domain-related knowledge made it difficult for novices to describe their cinematographic aspects and that this type of tags would be more used by experts. Unexpectedly, novices also used this type of tags, but in a more general fashion than experts did (for instance, as shown in **Table 2,** by using tags such as 'black and white', or 'silent film'). In relation to our first question, about how experts and novices' tags differ, **Table 4** confirms an important distinction, which is the experts' variety of domain-specific terms in relation to cinematographic language. These terms are located in the long-tail portion of the expert tags' distribution and are thus not quantitatively significant, but semantically rich from a qualitative perspective.

Cinematographic tags (sub-type)	**Expert tags' frequencies (Groups A+B)**	**Novice tags' frequencies (Groups C+D)**
Acting	extras (1); silent film actress (1)	
Copy	restoration (1);	poor picture quality (1)
Editing	rapid cutting (1); parallel cutting (1); reverse editing (1); continuity editing (1);	continuous (1) ; fadeout (1)
Genre	silent film *(mute cinema, mute pictures, silent, silent cinema, silent movie, silent movies)* (25); fiction (4); thriller (3); sound film (2); trailer (2); horror (2); drama (2); documentary feel (1); science fiction (1); melodrama (1)	silent film *(mute cinema, mute pictures, silent, silent cinema, silent movie, silent movies)* (25); fiction (1); thriller (1); horror (1); drama (3)
Mise-en-scene	exterior shots (3); interior shot *(interior scene)* (3); interior (2); decor (1); set design (1); setting (1)	
Narrative	intertitle (7); titles (4); credits (4); intro (2); climax (2); German intertitles (1); end title (1); title card (1); epilogue (1); narrative (1); end (1)	titles (1); end (2); start (1); subtitles (1); sequence (1)
Shot type-framing	close-up (6); long shot (4); high angle (3); camera pan (2); subjective shot (2); shot on location (1); pan shot (1); fear in close-up shot (1); deep focus (1); detail (1); diagonal (1); panning (1); point-of-view (1); crane shot (1); close up interior shots (1); offscreen (1); extreme long shot (1); topshot (1); low angle (1); aerial shot (1)	close-up (5)
Shot-photographic aspects	black-and-white film *(black and white, black & white, black white)* (10); superimposition (3); shadow theatre *(chinese shadows, javanese shadows, shadowplay)* (3); chiaroscuro (1); double exposure (1); vignetting on film (1); tableau (1); trick photography (1);	black-and-white film *(black and white, black & white, black white)* (22); shadow theatre *(chinese shadows, javanese shadows, shadowplay)* (1)

	silhuoettes (1); masking (1)	
Technique-sound	offscreen sound (2); scored music (1); accompaniment (1); musical accompaniment (1)	

Table 4. Cinematographic tags for the five film clips used by experts and novices groups combined (respectively A+B; C+D), including tags in the long-tail portion of the total tags' distribution.

We explored the semantic overlap of this tags' sub-set with The International Federation of Film Archives (FIAF) thesaurus (17), looking for similarity (syntactic and semantic) between the sample of tags in **Table 4** and the thesaurus descriptors. From the 77 Cinematography tags, only 10% (n=8) had an exact equivalent (syntactic and semantic); 32% (n=25) had some equivalent in the thesaurus (e.g. for the tag 'silent film' the correponding term would be "history of cinema. silent period"). None of the tags indicating shot type were found in the thesaurus, where the broader terms "Camera angles" or "Cinematography" cover all the spectrum. Other indexing alternatives different to thesaurus-based indexing are extensively investigated by Turner (e.g., 2009).

We assume that there are richer semantic connections within the tags themselves, and not only in relation to external vocabularies that do not have a time-based focus. In this sense, a relevant topic for future work is mining the semantic associations between tags and tag provenance in relation to the time dimension. For example, within a 10-second span, we can have a combination of expert and novice tags such as 'abandoned', 'house', 'panning'. If the tag 'panning' was added by a film expert, this could eventually indicate that there is a pan shot of an abandoned house in that time frame.

From **Table 3**, there does not seem to be any significant difference between the groups in the use of the tags of the type **"Explanatory"**. These tags range from the simple registry of objects and actions, to the higher level of abstract ideas, symbolic interpretations or interconnections (for instance, finding a relation with an art or literary movement, as in the tag 'expressionism'). These tags require from the tagger more effort in using her/his background knowledge, whether film related or not. In our test, both film experts and novices provided this type of tags to a low extent.

The **"Other"** category also lacks a significant difference. These tags mostly correspond to what in Classification No.2 is categorized as "Non-visual" level. It covers descriptive metadata such as the date (e.g. '1912', '1932'), location or country of origin ('french movie' '), creator (e.g. 'Dreyer'), title ('metropolis'), or historical-contextual aspects (e.g. 'early cinema').

Following the procedure used in Gligorov et al. (2011), we used Classification No.2 to filter out only the “conceptual” tags for the subsequent Panofsky-Shatford analysis (Classifications No.3 and 4). Tags classified in this category (“conceptual”) corresponded to 86% of the tags’ total (coincidentally this proportion is almost the same one found by Hollink (2006), who concluded in her empirical study about the use of the different categories in her model –our Classification No.2- that the conceptual levels were used most (87%)). **Table 5** shows the proportions of “conceptual” tags for the most frequent Panofsky/Shatford categories.

Category / Group	A Experts /no instructions	B Experts /instructions	C Novices /no instructions	D Novices /instructions	Total
General/Who (e.g., man, bell, dog, animals)	48.16%	40.27%	35.64%	32.59%	38.54%
General/What (e.g., bell ringing, children playing, hug, kissing goodbye)	23.21%	23.03%	21.19%	31.07%	24.88%
Abstract/What (e.g., abandoned, bored, calamity, danger)	15.09%	23.33%	26.37%	27.60%	23.63%
Abstract/Who (e.g., thief, proletarian, friend)	4.84%	7.73%	8.95%	4.99%	6.67%

Table 5. Proportional distribution of Conceptual tags across different categories (Classifications No.3 and 4: the Panofsky/ Shatford matrix) per group. Percentage in relation to the total conceptual tags per group.

In relation to RQ1, the figures in **Table 5** confirm our previous finding of the lack of substantial dissimilarities in the most common semantic types of tags by both groups. In this case, both experts and novices used more tags of the type “General/Who”, with no significant statistical difference between groups. This category corresponds mostly to factual tags and more specifically, to descriptions of objects in the scenes. This result agrees with Thøgersen (2013) who found in his study about still image tagging by general users that most tags were of the type “Artifact/objects”. After this category, tags in the “General/What” category predominate; these are descriptions of what happens in the scenes at a general level (e.g. ‘bell ringing’).

“Abstract/What” tags were the third more used type by both experts and novices, which corresponds to descriptions of events or actions in the scenes at an abstract level (e.g. ‘calamity’). In this category there was a statistically significant difference between groups A and C: **Table 5** shows that non-instructed novices (group C) tended to use more “abstract/what” tags than non-instructed experts (group A) (26.37% vs 15.09% respectively; *p=0.031* after a Mann–Whitney *U* test). These tags coincide with explanatory and emotional tags, which are of a more abstract nature.

Category/Group	A Experts /no instructions	B Experts /instructions	C Novices /no instructions	D Novices /instructions	Total
General	74.76%	66.72%	59.18%	64.63%	65.88%
Abstract	21.19%	31.51%	36.58%	34.40%	31.49%
Specific	4.05%	1.78%	4.24%	0.97%	2.62%

Table 6. Proportional distribution of Conceptual tags across different categories (Classification No.3) per group. Percentage in relation to the total conceptual tags per group.

In relation to RQ2, about the effect of instructions in the tags' selection, we found that instructed experts (group B) tended to use more abstract terms than their counterpart group without instructions (group A) (*p= 0.040,* from a U Mann-Whitney Test for groups A and B in the abstract category using Classification No.3*)*. This difference was due to the increased use of "Abstract/Who" tags by the instructed expert group (B) in relation to the non-instructed expert group (A) (*p=0.031,* using values from **Table 5***)*. The experts' preference for general tags over abstract tags (Table 6) shows similarities with conclusions reached by Thom-Santelli, Cosley, & Gay (2010). In their study about the differences between experts and novices in a collaborative environment, they found that experts have a preference for objective tags. The preference for this type of tags in a video labeling game also agrees with Gligorov et al. (2011), who found that most conceptual tags were general (74%). In our test, percentages of "abstract" tags were higher (31% of the total conceptual tags) than in Gligorov's study (7% of the total conceptual tags). This difference may be caused both by the type of content (film in our study vs. television in their study) and/or by the guidelines given to the taggers, which included "Emotions" in the possibilities.

Perception of the value of instructions

Participants in the guided groups (B and D) were positive about the value of instructions in helping them to come up with tags. **Table 7** shows that when asked about the value of the given instructions (q18) (18), the median from groups B and D is higher than for the non-instructed groups (A and C). A higher value of instructions was perceived among the novices group (D).

Groups (n=9)	q18.Perceived usefulness of instructions (categories) (1=not at all useful; 5=extremely useful)

	Mode	Median	Min	Max
Group A (Experts/no instructions)	2 (n=4)	2	1	5
Group B (Experts/instructions)	3 (n=3) 5 (n=3)	4	1	5
Group C (Novices/no instructions)	3 (n=6)	3	1	5
Group D (Novices/instructions)	3 (n=4) 4 (n=4)	4	3	5

Table 7. Frequencies of ranking on a 5 point Likert scale the usefulness of instructions during tagging (1=not at all; 5=extremely).

A number of non-instructed experts and novices (n=5) suggested that the categories that we used in the questionnaire to ask them rank the types of tags they used (Facts, Emotions, etc.) (q16) could have been used in the instructional text as guidance for which types to use. These reactions indicate that instructions about types of tags are necessary for time-based tagging. Participants described in the open answers to the questionnaire several issues which can be summarized in these points: (a) taggers need to know which aspects or dimensions they should focus on during tagging; presenting several types of tags in the instructions may help, but the participant needs only one to keep the focus; (b) participants should have previous knowledge about the movies and clips (e.g., contextual or historical information, and information about the clip itself), (c) the future retrieval purpose of the tagging activity should be stated; and (d) term suggestions may help the tagger.

The role of professional experience with indexing, tagging and labeling games

Lee, Goh, Razikin, & Chua (2009) showed that “the familiarity of users with the concept of tagging, the functionality of tagging systems, and the use of web catalogs has a great effect on the user’s tagging behavior” (p.184). To observe these issues, we asked the participants to rate their level of professional experience with indexing/cataloguing, their familiarity with creating tags, words or keywords for online content (for example: labeling images in Flickr, or videos in Youtube, or bookmarks in Delicious); about their familiarity level with video search through keywords or tags, and their knowledge and experience with video labeling games.

However, we did not find a statistically positive correlation between the number of tags entered by the participants and each one of these different aspects (using the Spearman’s Rho two-tailed test). This may be attributed to the quite homogenous “indexing” expertise of our participants regardless of their domain expertise. This leads us to be cautious about concluding that our study contradicts results from Lee et al. (2009), but rather that testing mechanisms for tagging familiarity should be refined in future tests.

The influence of content, and familiarity with the content

As expected, the domain expert participants reported familiarity with some of the video clips: (1) with “Metropolis” (n=15 high familiarity, and n=3 medium familiarity), “Vampyr” (n=7 high familiarity, and n=4 medium familiarity); and “Den flyvende circus” (n=1 high familiarity, and n=3 medium familiarity). There was a positive statistical correlation between the most familiar clip for all participants (“Metropolis”) and its total number of tags (*r=0.442; p=0.007 from a* Spearman’s Rho two-tailed test for testing correlation between familiarity with each film and its corresponding number of tags for this clip), which indicates that a higher level of familiarity resulted in more tags. There is also a negative correlation between familiarity with this film and the use of emotional tags (*r=-0.461; p=0.005)*, which indicates that the more familiar the tagger was with this film, the less likely was to use emotional tags. This corresponds to our previous findings of a marginal significant difference in the number of tags at the video level for the clip of “Metropolis”. In this case, the experts’ groups entered more tags than the novices’ groups, and those tags were significantly less of the type “Emotions”. Instead, “Explanations” and the “Other” tags’ categories were more used, which reflect experts’ knowledge about the metadata attributes and interpretations of this movie (e.g., ‘dystopia’, ‘Fritz Lang’).

From the participants’ answers, it was also observed that familiarity with the content plays an important role in motivating the tagger. It also allows the participant to concentrate on tagging, and not on getting acquainted with a movie that is new for her/him. As one expert states: *“there is always the difference between knowing a film and seeing it for the first time. The first time [you have] reactions on what you see, the second time is more intentional” (Participant group B).*

Game effect, scoring and tagging motivations

A common feeling among the participants from all groups was time pressure. They found that the short duration of the clips, or the impossibility to replay them, added stress to think of, or limited them to entering more tags, both during the video (because they were watching it and not entering tags) or at the end of the clip (tags for the last frames). One expert commented that this was not “a professional way of working” *(Participant group A)*.

Groups (n=9)	q12.Difficulty in coming up with tags				q13.Possibility of entering all tags				q15.Influence of scoring in game motivation			
	Mode	Median	Min	Max	Mode	Median	Min	Max	Mode	Median	Min	Max
Group A (Experts/no instructions)	2 (n=3); 4 (n=3)	4	2	5	4 (n=5)	4	1	5	1 (n=3) 2 (n=3)	2	1	5
Group B (Experts/instructions)	3 (n=3) 4 (n=3)	3	2	5	2 (n=3) 3 (n=3)	3	2	5	4 (n=3)	3	1	5
Group C (Novices/no instructions)	2 (n=4)	3	2	5	4 (n=4)	4	3	5	1; 2; 4; 5 (n=2)	3	1	5
Group D (Novices/instructions)	2 (n=3); 3 (n=3)	3	2	5	4 (n=4)	4	1	5	4 (n=3)	4	1	5

Table 8. Frequencies of ranking on a 5 point Likert scale different aspects of tagging behavior: (q12: 1=very difficult; 5=very easy); (q13: 1=not possible; 5= possible); (q15: 1=not at all influential; 5=extremely influential).

From **Table 8,** we can conclude that it seemed to be easier for the experts groups (A+B) to come up with tags than for the novices. Among the instructed experts group (B) there were participants dissatisfied for not being able to enter all tags that occurred to them. They explained that the lack of familiarity and short duration challenged them in this respect. Participants from different groups pointed to different negative issues related to the game influence. These include (a) "multitasking" (i.e. watching the video, thinking of tags, typing it in); typing skills (having to look at the keyboard); (b) the impossibility to synthesize in a single word or in a couple of words the concepts they had about the fragments, and/or to recall the technical terms referring to shot types and editing; (c) language issues and spelling.

The reaction to scoring and gaming elements (q15) are very personal, and we cannot conclude any relation to domain expertise. Some experts made positive comments about the game itself and found it fun. Both among the experts and novices groups there were few participants concerned for having few matching tags. Not surprisingly, we found a positive correlation between scoring motivation and number of tags (*r=0.406, p=0.014* after a Spearman's Rho two-tailed test). A drawback of this correlation, also identified by Thøgersen (2013), is that since the game is set up to reward players based on matching tags, this encourages most players to tag what is in the picture, rather than thinking about other possibilities.

Finally, as in other tagging activities, there should be a quality control and feedback mechanism that allows the participant to check the value of her/his tags. One novice said: *"It was very easy to write a tag when it came up in mind. The only difficulty was in deciding if it was a "correct" tag, i.e. if the word made sense or it was just an instinctive reaction to what I was seeing" (Participant group C).*

We can conclude that clear guidance and objectives in the tagging activity, encouraging participants to use their specific domain knowledge, and a flexible tagging setting (not necessarily competitive), may increase the motivation in the tagging activity beyond scoring mechanisms. Future work should focus on investigating which rewarding mechanisms work better for experts. One direction is suggested in the study by Thom-Santelli et al. (2010), who points to innate experts' feelings of territoriality and "curation", which means that experts can have higher levels of participation due to ownership feelings in cooperative work that involves targets of their concern (e.g. museum objects).

Perception of the utility of selected tags

Table 9 indicates that novices were more positive about the possible use of their tags for future retrieval of the videos than experts, who were mostly uncertain.

Groups (n=9)	**q20.Perceived usefulness of entered tags** (No=0 / Uncertain=1 /Yes=2)	
	Mode	**Median**
Group A (Experts/no instructions)	1 (n=4)	1
Group B (Experts/instructions)	1 (n=6)	1
Group C (Novices/no instructions)	2 (n=8)	2
Group D (Novices/instructions)	2 (n=6)	2

Table 9. Frequencies of ranking of perceived utility of each participant's contributed tags.

Indeed, domain experts cast doubt on the tags' semantic value. They consider them very general and only related to describing what they saw in the images, without taking into account any context. For these experts, this does not correspond to describing the actual content of the film. For instance, one expert stated: *"My tags were very factual, about what you see in the image. If you want footage of a train, then you will find L'aiguille. If you are looking for a silent expressionist horror film, you will not find Vampyr with my tags" (Participant group A)*. One more expert confirms the utility of her/his tags, but, as s(he) says: *"only for such purposes as stock video footage, but not for meeting thematic or content driven curatorial or research needs" (Participant group A).*

This shows the need for more research in understanding the use of time-based annotations for research purposes, beyond footage finding. From the novices perspective there are other concerns, one novice commented: *"I guess moviegoers tend to select films based on the genre as well as actors/actresses and maybe directors involved with the film. I am wondering how social tagging plays a part in helping us decide which films to watch" (Participant group D).* Current practice is showing interesting directions in involving humans in creating keywords for movie recommendation for entertainment, such as the Netflix case described by Madrigal (2014). These practices have roots in cultural heritage curation, and film archives can benefit from them for dissemination purposes.

Limitations

The data collection took place in a game setting, which may be a very specific type of tagging scenario. However, even though this study did not include a comparison between the differences with non-game contexts, most of the findings were in line with conclusions found in other experiments based on other data collection methods.

In relation to homogeneity in the experts and novices groups, we did not include in our procedures a detection and/or operationalization of expertise by testing the actual knowledge of the participants (as it is done for instance in Kang & Fu, 2010). Additionally, we omitted any form of control in the participants who got the instructions to know if they read them in detail; at least one participant admitted to having skipped a careful reading.

About the labeling setting, we chose to let participants play against a bot, instead of the default setting: against each other. Influence in tag selection by the participants is, in both cases unavoidable and difficult to judge or measure.

We find the procedure of allowing taggers to use their mother tongue valid for our research purposes, but multilingualism is far from being a trivial issue, and a research area on its own that we did not touch in depth it in our study.

Finally, this was a small-scale experiment that counted with the participation of the minimum number of film experts and novices (45 cases per group: 5 videos x 9 participants). A higher number of participants would be needed to validate the findings quantitatively.

Conclusions and future work

Coming back to RQ1, about how domain experts tag film content in comparison to novices, we observe that experts tag in a similar fashion as novices when participating in a tagging game. In general they enter the same number of tags, and they mostly use "Factual" tags. However, in the experts' less-frequent tags, we found more domain-specific terminology than in the novices groups.

The use of the most common type of tags ("Facts") among the two groups, agrees with other studies on image subject categorization (Klavans, LaPlante, & Golbeck, 2013), with other game related experiments (Thøgersen, 2013), and with the tag analysis of the first *Waisda?* projects for TV broadcasts. These tags describe the content at a general level (Gligorov et al., 2011). Perhaps, as Halpin et al. (2007) indicate, tagging requires less cognitive effort, which would explain why experts tagging behavior was similar to the one of novices, at least under the conditions given the chosen labeling setting. And yet we think that a clearer explanation for the groups' similarity is the competitive nature of the game. We confirm this is not the best scenario to tap into the domain-specific knowledge of experts (as it was somehow expected, and also pointed out by the experts themselves in their comments). However, the same game has proved to be useful for getting a relatively high number of relatively high quality time-stamped tags from general users (as Ahn & Dabbish, 2008; Gligorov et al., 2013 found out). This poses the issue of how to join the advantages of a great number of common tags (which can improve indexing consistency, assumed to indicate quality (Good et al., 2009, p. 6)) with less frequent expert tags, assumed to be more relevant for specialized contexts (Tsai et al., 2011). In this regard, we confirm the need for extracting tag provenance information, which can add to the quality measures of the tags. This follows the tendency to mining not only the relationships between tags and documents, but the link between users, tags, and documents (as suggested by Good et al., 2009). In addition to this, one aspect that was not possible to cover in this study, but which needs future exploration is the analysis of the influence of film genre in the types of tags.

In what concerns RQ2, about the influence of guidelines in the selection of types of tags, we conclude that novices can provide tags in different types of categories. However, as expected, the level of detail of the individual tags in the most domain-related category (Cinematography) is limited. Additionally, from the questionnaire we know that most participants preferred to have a clear description of the type of tags they were expected to enter. In the case of moving images, where several dimensions co-occur, instructions should help participant focus on specific content or stylistic aspects and allow complementarity of novice and expert tags for the same video. For instance, one of the usage scenarios for online film archives to enrich and give access to their online digital collections could be to ask experts to contribute only “Cinematography” tags. In this way, film experts’ tags could be used for novices in browsing and learning the cinematographic language because expert tags seem to have the potential to augment the exploratory search of information. This holds especially for users who have little knowledge on a topic (as Kang & Fu (2010) found). Novices, on the other hand, should be guided to contribute “Facts” (and eventually emotions or explanations) in their tags, according to expertise in other domains, not necessarily film-related backgrounds. In any case, *nichesourcing* initiatives are about channeling expert knowledge instead of asking experts to do what novices, or eventually content-based retrieval algorithms, could also do.

More studies need to be done to understand the way of motivating and obtaining significant time-based tags or annotations from film experts and novices for research or educational purposes, and not only for footage finding. Current research also points to the fact that tagging implementations should be part of more integrated curatorial and annotation infrastructures, and that isolated tagging support may not be the best solution to obtain expert tags. One example is “The Larm Project” (Skov & Lykke, 2012) in which a national research infrastructure for radio and audio-based research is built through a collaboration between universities and radio archives. This infrastructure would support knowledge dissemination, sharing and interaction between different kinds of humanities researchers, by providing necessary scholarly-based links between texts and images (Winget, 2009). In fact, in these settings, games are still an option, though not the only one. A requirement is that more varied genres of a higher collaborative nature are investigated, as pointed out in Goh et al. (2011).

We are currently exploring if we can use expert descriptions made outside the game setting to improve the tags made by novices inside the game. Also, we are exploring the use of term suggestions from, for example, the IMDB plot keywords (19) database, or from technical film glossaries. Although these techniques are already in use, more theoretical work needs to be done to provide semantic models and classifications schemes specific for moving images.

Acknowledgements

This study was possible thanks to the enthusiastic participation of 36 experts and novices who played the

Waisda? game and contributed with insightful comments. We also thank Erwin Verbruggen and Hugo Huurdeman for contacts and advice, and our reviewers for the valuable suggestions.

Footnotes

(1). http://www.bbc.co.uk/radio4/science/findlistenlabel/

(2). http://celluloidremix.openbeelden.nl/

(3). http://www.scenemachine.nl/

(4). http://tagger.thepcf.org.uk

(5). http://prestoprime.cs.vu.nl/efg

(6). https://github.com/beeldengeluid/waisda

(7). http://www.europeanfilmgateway.eu/content/about-european-film-gateway

(8). “Den Flyvende Cirkus” at EFG: http://tinyurl.com/p8cutp5.

Film by the Film Fabrikken Danmark production company. Directed by Alfred Lind (1879-1959), whose name is “inextricably linked with a large part of Danish silent film milestones”, according to the Danish National Filmography (http://www.dfi.dk/faktaomfilm/person/da/127597.aspx?id=127597).

(9). “Die Gezeichneten” at EFG: http://tinyurl.com/nhrdpn6

Original title “Elsker hverandre” (Love one another). Directed by Carl Theodor Dreyer, recognized to be Danish cinema’s most important director; (Danish National Filmography, http://www.dfi.dk/faktaomfilm/person/da/7401.aspx?id=7401).

(10). “L’aiguille” at EFG: http://tinyurl.com/l9yp4qg

Original title: “Die Weiche”. Swiss short feature film produced in 1961. It is an unknown film from an unknown director. The EFG portal does not give detailed contextual information about it. Some film scholars think it is an amateur film, which combines different cinematographic techniques in naïve approach too basic for its time (November, 2014, personal communication with Spanish film scholars).

(11). “Metropolis” at EFG: http://tinyurl.com/kmvmylh

Fritz Lang's classic and renowned science fiction film, one of the greatest films of all times. The clip corresponds to the sequence where the robot Maria incites the workers to revolt.

(12). “Vampyr” at EFG: http://tinyurl.com/otunuvv

Also known as “L’etrange aventure de David Gray”, is one of the most known films by Carl Theodor Dreyer and is “one of the founding and defining works of psychological horror cinema” (Rudkin, 2007).

(13). The questionnaire is available at surveys.timelessfuture.com/waisda

(14). The data is made available online in anonymized form at https://github.com/biktorrr/waisda_efg

(15). The “Emolab” project by Frans Hals Museum in Haarlem, The Netherlands. http://www.commit-nl.nl/news/emolab-in-frans-hals-museum

(16). Project “Emotions in Film” at the University of Amsterdam. http://cdh.uva.nl/projects-2012-2013/emotions-in-film/emoties-in-film.html

(17). http://www.fiafnet.org/uk/publications/iifp_subjectHeadings.html

(18). Questions are numbered “q1, q2,…” the full questions are found in the Appendix.

(19). http://www.imdb.com/Sections/Keywords/

References

(All online documents were last accessed on January 26, 2015).

Ådland, M. K., & Lykke, M. (2012). Social Tagging in Support of Cancer Patients' Information Interaction. In *Social Information Research* (Vol. 5, pp. 101–128). UK: Emerald Group Publishing.

Ahn, L. von, & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67. http://doi.org/10.1145/1378704.1378719

Baca, M. (2002). *Introduction to Art Image Access: Issues, Tools, Standards, and Strategies*. Los Angeles, CA: Getty Research Institute.

Bálint, K., & Kovács, A. B. (2012). Focalization and Attachment. Studying the interaction effect of narrative and psychological factors in film viewers' emotional responses. *Pszichológia*, *32*(3), 271–291. http://doi.org/10.1556/Pszicho.32.2012.3.6

Ballan, L., Bertini, M., Del Bimbo, A., Meoni, M., & Serra, G. (2010). Tag Suggestion and Localization in User-generated Videos Based on Social Knowledge. In *Proceedings of Second ACM SIGMM Workshop on Social Media* (pp. 3–8). New York, NY, USA: ACM. http://doi.org/10.1145/1878151.1878155

Ballan, L., Bertini, M., Del Bimbo, A., & Serra, G. (2011). Enriching and Localizing Semantic Tags in Internet Videos. In *Proceedings of the 19th ACM International Conference on Multimedia* (pp. 1541–1544). New York, NY, USA: ACM. http://doi.org/10.1145/2072298.2072060

Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y., & Shachak, A. (2008). Structured versus unstructured tagging: a case study. *Online Information Review*, *32*(5), 635–647. http://doi.org/http://dx.doi.org/10.1108/14684520810914016

Bertini, M., Del Bimbo, A., Ferracani, A., Gelli, F., Maddaluno, D., & Pezzatini, D. (2013a). A Novel Framework for Collaborative Video Recommendation, Interest Discovery and Friendship Suggestion Based on Semantic Profiling. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 451–452). New York, NY, USA: ACM. http://doi.org/10.1145/2502081.2502264

Bertini, M., Del Bimbo, A., Ferracani, A., Gelli, F., Maddaluno, D., & Pezzatini, D. (2013b). Socially-aware Video Recommendation Using Users' Profiles and Crowdsourced Annotations. In *Proceedings of the 2Nd International Workshop on Socially-aware Multimedia* (pp. 13–18). New York, NY, USA: ACM. http://doi.org/10.1145/2509916.2509924

Boer, V. de, Hildebrand, M., Aroyo, L., Leenheer, P. D., Dijkshoorn, C., Tesfa, B., & Schreiber, G. (2012). Nichesourcing: harnessing the power of crowds of experts. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, … N. Hernandez (Eds.), *Ekaw'12: proceedings of the 18th international conference on knowledge engineering and knowledge management* (pp. 16–20). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://doi.org/10.1007/978-3-642-33876-2_3

Bordwell, D., & Thompson, K. (2003). *Film Art: An Introduction* (7th ed.). Mcgraw-Hill College.

Burford, B., Briggs, P., & Eakins, J. P. (2003). A Taxonomy of the Image: On the Classification of Content for Image Retrieval. *Visual Communication*, *2*(2), 123–161. http://doi.org/10.1177/1470357203002002001

Darvish, S., & Chin, A. (2010). Dealing with the Video Tidal Wave: The Relevance of Expertise for Video Tagging. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (pp. 289–290). New York, NY, USA: ACM. http://doi.org/10.1145/1810617.1810679

Eakins, J. P., Briggs, P., & Burford, B. (2004). Image Retrieval Interfaces: A User Perspective. In P. G. B. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, & A. W. M. Smeulders (Eds.), *Image and Video Retrieval* (pp. 628–637). Springer Berlin Heidelberg.

Enser, P. G. B. (2000). Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science*, *26*(4), 199–210. http://doi.org/10.1177/016555150002600401

Fossati, G. (2009). *From Grain to Pixel: The Archival Life of Film in Transition*. Amsterdam University Press.

Freiburg, B., Kamps, J., & Snoek, C. G. M. (2011). Crowdsourcing visual detectors for video search (pp. 913–916). New York, NY, USA: ACM. http://doi.org/10.1145/2072298.2071901

Fu, W.-T., Kannampallil, T., Kang, R., & He, J. (2010). Semantic imitation in social tagging. *ACM Trans. Comput.-Hum. Interact.*, *17*(3), 12:1–12:37. http://doi.org/10.1145/1806923.1806926

Gedikli, F., & Jannach, D. (2013). Improving Recommendation Accuracy Based on Item-specific Tag Preferences. *ACM Trans. Intell. Syst. Technol.*, *4*(1), 11:1–11:19. http://doi.org/10.1145/2414425.2414436

Geisler, G., Willard, G., & Whitworth, E. (2010). Crowdsourcing the indexing of film and television media (pp. 82:1–82:10). Silver Springs, MD, USA: American Society for Information Science. Retrieved from http://dl.acm.org/citation.cfm?id=1920331.1920448

Gibbon, D. C., Liu, Z., Basso, A., & Shahraray, B. (2013). Automated content metadata extraction services based on MPEG standards. *Computer Journal*, *56*(5), 628–645. http://doi.org/10.1093/comjnl/bxs146

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Aroyo, L., & Schreiber, G. (2013). An Evaluation of Labelling-Game Data for Video Retrieval. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, … E. Yilmaz (Eds.), *Advances in Information Retrieval* (pp. 50–61). Springer Berlin Heidelberg.

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Schreiber, G., & Aroyo, L. (2011). On the role of user-generated metadata in audio visual collections (pp. 145–152). New York, NY, USA: ACM. http://doi.org/10.1145/1999676.1999702

Goh, D. H.-L., Ang, R. P., Lee, C. S., & Chua, A. Y. K. (2011). Fight or unite: Investigating game genres for image tagging. *Journal of the American Society for Information Science and Technology*, *62*(7), 1311–1324. http://doi.org/10.1002/asi.21478

Goh, D. H.-L., & Lee, C. S. (2011). Perceptions, quality and motivational needs in image tagging human computation games. *Journal of Information Science*, *37*(5), 515–531. http://doi.org/10.1177/0165551511417786

Golbeck, J., Koepfler, J., & Emmerling, B. (2011). An experimental study of social tagging behavior and image content. *Journal of the American Society for Information Science and Technology*, *62*(9), 1750–1760. http://doi.org/10.1002/asi.21522

Good, B. M., Tennis, J. T., & Wilkinson, M. D. (2009). Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics. *BMC Bioinformatics*, *10*, 313–313. http://doi.org/10.1186/1471-2105-10-313

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up Tags? *D-Lib Magazine*, *12*(1). Retrieved from http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january06/guy/01guy.html

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web* (pp. 211–220). New York, NY, USA: ACM. http://doi.org/10.1145/1242572.1242602

Hildebrand, M., Brinkerink, M., Gligorov, R., van Steenbergen, M., Huijkman, J., & Oomen, J. (2013). Waisda? Video Labeling Game. Presented at the ACM Multimedia, Barcelona.

Hollink, L. (2006). *Semantic annotation for retrieval of visual resources* (Doctoral Dissertation). Vrije Universiteit, Amsterdam. Retrieved from http://hdl.handle.net/1871/10846

Hollink, L., Schreiber, A. T., Wielinga, B. J., & Worring, M. (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies*, *61*(5), 601–626. http://doi.org/10.1016/j.ijhcs.2004.03.002

Huang, C., Fu, T., & Chen, H. (2010). Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, *61*(5), 891–906. http://doi.org/10.1002/asi.21291

Images for the Future. (2009). Waisda? Video Labeling Game: Evaluation Report. Retrieved from http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

Inskip, C., MacFarlane, A., & Rafferty, P. (2008). Content or Context?: Searching for Musical Meaning in Task-based Interactive Information Retrieval. In *Proceedings of the Second International Symposium on Information Interaction in Context* (pp. 72–74). New York, NY, USA: ACM. http://doi.org/10.1145/1414694.1414711

Kang, R., & Fu, W.-T. (2010). Exploratory information search by domain experts and novices. In *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 329–332). New York, NY, USA: ACM. http://doi.org/10.1145/1719970.1720023

Kim, H. H., & Kim, Y. H. (2010). Toward a conceptual framework of key-frame extraction and storyboard display for video summarization. *Journal of the American Society for Information Science and Technology*, *61*(5), 927–939. http://doi.org/10.1002/asi.21317

Klavans, J. L., LaPlante, R., & Golbeck, J. (2013). Subject matter categorization of tags applied to digital images from art museums. *Journal of the American Society for Information Science and Technology*, n/a–n/a. http://doi.org/10.1002/asi.22950

Layne, S. S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging & Classification Quarterly*, *6*(3), 39–62. http://doi.org/10.1300/J104v06n03_04

Lee, C. S., Goh, D. H.-L., Razikin, K., & Chua, A. Y. K. (2009). Tagging, Sharing and the Influence of Personal Experience. *Journal of Digital Information*, *10*(1). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/275

Li, G., Wang, M., Zheng, Y.-T., Li, H., Zha, Z.-J., & Chua, T.-S. (2011). ShotTagger: Tag Location for Internet Videos. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (pp. 37:1–37:8). New York, NY, USA: ACM. http://doi.org/10.1145/1991996.1992033

Lu, C., Park, J., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, *36*(6), 763–779. http://doi.org/10.1177/0165551510386173

Madrigal, A. C. (2014, January). How Netflix Reverse Engineered Hollywood. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/3/

Matusiak, K. K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services*, *22*(4), 283–298. http://doi.org/10.1108/10650750610706998

Melenhorst, M., Grootveld, M., van Setten, M., & Veenstra, M. (2008). Tag-based information retrieval of video content (pp. 31–40). New York, NY, USA: ACM. http://doi.org/10.1145/1453805.1453813

Oomen, J., & Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges (pp. 138–149). New York, NY, USA: ACM. http://doi.org/10.1145/2103354.2103373

Panofsky, E. (1939). *Studies in iconology : Humanistic themes in the art of the Renaissance* (1st Icon ed., 4th print). New York, NY: Harper & Row.

Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0* (1st ed.). De Gruyter.

Rudkin, D. (2007). *Vampyr*. London: University of California Press.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., … Riedl, J. (2006). Tagging, Communities, Vocabulary, Evolution. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (pp. 181–190). New York, NY, USA: ACM. http://doi.org/10.1145/1180875.1180904

Skov, M., & Lykke, M. (2012). Unlocking radio broadcasts: user needs in sound retrieval (pp. 298–301). New York, NY, USA: ACM. http://doi.org/10.1145/2362724.2362779

Smith, G. (2007). *Tagging: People-powered Metadata for the Social Web*. New Riders Press.

Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D., & Zinkham, H. (2008). *For the* *common good: the Library of Congress Flickr pilot project* (Report). D.C.: Government of the United States; Library of Congress. Retrieved from http://www.egov.vic.gov.au/focus-on-countries/north-and-south-america-and-the-caribbean/united-states/government-initiatives-united-states/culture-sport-and-recreation-united-states/libraries-united-states/for-the-common-good-the-library-of-congress-flickr-pilot-project-in-pdf-format-1333kb-.html

Thøgersen, R. (2013). Data Quality in an Output-Agreement Game: A Comparison between Game-Generated Tags and Professional Descriptors. In P. Antunes, M. A. Gerosa, A. Sylvester, J. Vassileva, & G.-J. de Vreede (Eds.), *Collaboration and Technology* (pp. 126–142). Springer Berlin Heidelberg.

Thom-Santelli, J., Cosley, D., & Gay, G. (2010). What do you know?: experts, novices and territoriality in collaborative systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1685–1694). New York, NY, USA: ACM. http://doi.org/10.1145/1753326.1753578

Tirilly, P., Mu, X., Huang, C., Xie, I., Jeong, W., & Zhang, J. (2012). On the consistency and features of image similarity (pp. 164–173). New York, NY, USA: ACM. http://doi.org/10.1145/2362724.2362754

Trant, J. (2009a). Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing Research. *Journal of Digital Information*, *10*(1). Retrieved from http://journals.tdl.org/jodi/article/viewArticle/270

Trant, J. (2009b). *Tagging, Folksonomy and Art Museums: Results of steve.museum's research*. Retrieved from http://www.museumsandtheweb.com/blog/jtrant/stevemuseum_research_report_available_tagging_fo.html

Troncy, R., Huet, B., & Schenk, S. (2011). *Multimedia Semantics: Metadata, Analysis and Interaction*. John Wiley & Sons.

Tsai, L.-C., Hwang, S.-L., & Tang, K.-H. (2011). Analysis of keyword-based tagging behaviors of experts and novices. *Online Information Review*, *35*(2), 272–290. http://doi.org/http://dx.doi.org/10.1108/14684521111128041

Turner, J. M. (2009). Moving image indexing. In *Encyclopedia of Library and Information Sciences* (3rd ed., pp. 3671–3681). New York: Taylor & Francis.

Turner, J. M. (2010). From ABC to http: The Effervescent Evolution of Indexing for Audiovisual Materials. *Cataloging & Classification Quarterly*, *48*(1), 83–93. http://doi.org/10.1080/01639370903341919

Wang, M., Ni, B., Hua, X.-S., & Chua, T.-S. (2012). Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*, *44*(4), 25:1–25:24. http://doi.org/10.1145/2333112.2333120

Westman, S. (2009). Image Users' Needs and Searching Behaviour. In A. Göker & J. Davies (Eds.), *Information Retrieval: Searching in the 21st Century; Human Information Retrieval* (pp. 63–83). John Wiley & Sons, Ltd. Retrieved from http://doi/10.1002/9780470033647.ch4/summary

Wilkie, C. (1999). *Managing Film and Video Collections*. Aslib and Information Management International.

Winget, M. (2009). Describing art: an alternative approach to subject access and interpretation. *Journal of Documentation*, *65*(6), 958–976.

Yeh, M.-C., & Wu, W.-P. (2014). Clustering Faces in Movies Using an Automatically Constructed Social Network. *IEEE MultiMedia*, *21*(2), 22–31. http://doi.org/10.1109/MMUL.2014.24

Zollers, A. (2007). Emerging Motivations for Tagging: Expression, Performance, and Activism. In *Tagging and Metadata for Social Information Organization Workshop, WWW07*.

Time-based tags for fiction movies:

Comparing experts to novices using a video labeling game

Liliana Melgar Estrada*

Carlos III University of Madrid, Library and Information Science Department, Spain. E-mail: lmelgar@bib.uc3m.es

Michiel Hildebrand

Centrum Wiskunde & Informatica (CWI), The Netherlands. E-mail: Michiel.Hildebrand@cwi.nl

Victor de Boer**, and **Jacco van Ossenbruggen*

Vrije Universiteit Amsterdam, The Netherlands. E-mail: {v.de.boer, j.r.van.ossenbruggen} @vu.nl

*Corresponding author

** Jacco van Ossenbruggen is also affiliated with Centrum Wiskunde & Informatica (CWI).

Abstract. The cultural heritage sector has embraced social tagging as a way to increase both access to online content and to engage users with their digital collections. In this paper, we build on two current lines of research. (1) We use *Waisda?,* an existing labeling game, to add time-based annotations to content. (2) In this context, we investigate the role of experts in human-based computation (*nichesourcing*). We report on a small-scale experiment in which we applied *Waisda?* to content from film archives. We study the differences in the type of time-based tags between experts and novices for film clips in a *crowdsourcing* setting. The findings show high similarity in the number and type of tags (mostly factual). In the less frequent tags, however, experts used more domain-specific terms. We conclude that competitive games are not suited to elicit real expert-level descriptions. We also confirm that providing guidelines, based on conceptual frameworks that are more suited to moving images in a time-based fashion, could result in increasing the quality of the tags, thus allowing for creating more tag-based innovative services for online audiovisual heritage.

Keywords: social tagging, tagging games, nichesourcing, expert tags, audiovisual heritage, film, *Waisda?*

Introduction

In the cultural heritage domain, social tagging has become an attractive solution to involve the public in the process of describing the objects in digital collections (Oomen & Aroyo, 2011). For example, the Steve museum social tagging project collected a large number of tags that describe artworks (Trant, 2009a). The *Waisda?* video labeling game, launched in 2009 by the Netherlands Institute for Sound and Vision, was used in two projects to collect tags for TV broadcasts and historic newsreels, showing that social tagging can also be applied to the audiovisual domain (Gligorov, Hildebrand, van Ossenbruggen, Schreiber, & Aroyo, 2011; Images for the Future, 2009). Together, the two projects resulted in over a million time-based tags that describe the content in the video, for example, depicted locations.

Analysis of the tags collected with *Waisda?* for TV broadcasts showed that users primarily describe the visual content at a general level (Gligorov et al., 2011). Motion pictures, however, have a distinctive form and a specific narrative (Bordwell & Thompson, 2003, p. 2) and involve different semantic dimensions compared to TV broadcasts, such as the use of framing, camera movements and composition to express meaning. Tags at this specific level are needed to describe adequately and retrieve film content, for instance, when users do archival footage research, based on “shot listings” (Turner, 2010; Wilkie, 1999). It is unclear if players of a video labeling game would provide specific tags of this kind.

In this paper we investigate the difference in the types of tags provided by experts and novices with three aims: 1) contributing to the understanding of the role of expert tags for content access in the audiovisual heritage domain, in line with the studies on *nichesourcing*; 2) continuing research on time-based metadata and labeling games initiated by the *Waisda?* experiments, exploring to what extent a video labeling game can be used to collect tags for films; and 3) contributing to the overall discussion of how social tagging can be applied to the film domain. By *film domain,* we mean mostly fiction movies, not necessarily celluloid films.

For this purpose, we designed a small-scale experiment using *Waisda?*, in which both film experts and novices performed time-based tagging for five film clips. This study does not seek generalizations, but identification of emergent issues in social tagging and human computation research applied to film images.

First, we present prior work related to our study. We describe the experimental design and setting and report our results and discuss them. We present the limitations of this study, followed by the main conclusions and ideas for future work.

Related work

We discuss four main topics related to our study: social tagging in the audiovisual heritage domain, tags from experts versus novices, guided tagging, and tag categories and models for image description.

Social tagging in the audiovisual heritage domain

Social tagging has been one of the earliest implemented collaborative practices for describing shared content online. Since in 2005 services like Furl, Flickr, and Del.icio.us started offering their users the option to add labels or tags to organize content (Smith, 2007), many websites have incorporated social tagging services, and research has not ceased in discovering new theoretical and practical approaches to this way of indexing digital information.

The cultural heritage sector has embraced this practice and is progressively incorporating it, together with other *crowdsourcing* initiatives, as part of their workflows (Oomen & Aroyo, 2011). However, regarding access to audiovisual heritage through socially generated tags, research is just starting.

State-of-the-art automatic moving image retrieval can achieve content-based indexing based on the images' low-level features and concept-based indexing based on derived high-level concepts (Stock, 2010). However, the performance is still not optimal to be used in all settings (Gibbon, Liu, Basso, & Shahraray, 2013; Yeh & Wu, 2014). In turn, different techniques for semi-automatic concept-based indexing at the shot level have been investigated by Turner (2009) though they only apply at a small scale. But socially generated tags (by niche groups and by the general crowd), if well guided, could help to bridge the gap 1) between content-based and concept-based annotations (as promulgated by Enser, 2000; and explored in Freiburg, Kamps, & Snoek, 2011; and Melenhorst, Grootveld, van Setten, & Veenstra, 2008) and 2) among concept-based annotations created manually (as different studies with tags have shown, such as Lu, Park, & Hu, 2010; Matusiak, 2006; and Springer et al., 2008).

In the audiovisual domain, social tagging research has focused mainly on recommendations of entire videos or movies based on tags and user profiles (for instance in the work by Bertini et al., 2013a, 2013b, and Gedikli & Jannach, 2013), and in video classification based on tags (for instance in Huang, Fu, & Chen, 2010). Little research exists, however, about the application of tags to time-based metadata, also called "time-coded metadata", or "strata" by Troncy, Huet, & Schenk (2011, p. 7), which is the information related to a specific time frame within video sequences. This research gap has been identified in Ballan, Bertini, Del Bimbo, Meoni, & Serra (2010; 2011); and Li et al. (2011), even though on a practical side, initial implementations of time-based social tagging are emerging in the audio domain, for instance the BBC's "Find, listen, label" tool for adding notes to radio programs (1).

The few exceptions to the lack of research in this area include an early study about tagging applied to the movie recommendation service "MovieLens" (Sen et al., 2006). Most related to our work are studies related to a larger effort to develop a framework for the *crowdsourcing* of film and television indexing by Geisler, Willard, & Whitworth (2010). Other related work consists of a study by Freiburg, Kamps, & Snoek (2011), that looks at the time-based metadata approach in combination with socially generated tags and automatically created annotations to video fragments of music concerts; the Larm Project in the radiophonic cultural heritage which gives prominence to user-driven annotations (Skov & Lykke, 2012), and the studies done in the framework of the *Waisda?* project.

Waisda? is a social tagging application and research project in the audiovisual heritage domain. Specifically it uses the idea of games-with-a-purpose (Ahn & Dabbish, 2008) to motivate users to contribute, since play and competition have been identified as motivating factors for tagging (Zollers, 2007). It was launched in 2009 by the Netherlands Institute for Sound and Vision. During the first pilot, the site received more than 12,000 visits, and attracted over 2,000 players, contributing 420,000 tags for 604 video items (Gligorov et al., 2011; Images for the Future, 2009). In the second pilot approximately 750,000 tags were collected. This is in line with the increasing popularity of human computation games (HCGs) for image description (Goh, Ang, Lee, & Chua, 2011; Goh & Lee, 2011). HCGs are one way of harnessing human intelligence, through the use of computer games, to perform activities that are impossible to automate, such as distinguishing types of fruits in an image (Goh et al., 2011). The first *Waisda?* pilots showed that *crowdsourcing*, in the form of a labeling game, can be also a good way to engage audiences with collections while obtaining content descriptors that can enhance retrieval (Gligorov, Hildebrand, van Ossenbruggen, Aroyo, & Schreiber, 2013).

In the film domain, content keywords have been utilized successfully for Fossati calls the "creative re-use of, or inspiration by archival material" (Fossati, 2009, p. 96). Examples of this approach are the "Celluloid Remix contest" (2), and "The Scene Machine" (3), which allow users to creatively explore online archival film footage relying upon keyword-based search, and to use existing labels to create their own content. However, these keywords are not socially generated but provided by the coordinating institutions. They also do not seem to be based on research about generating social tags in a moving image context.

However, there is consensus in that socially generated tags have quality problems associated with the use of non-words, polysemy, synonymy and lack of hierarchy (Guy & Tonkin, 2006; Matusiak, 2006; Lu et al., 2010), and to the lack of distinction of which type a tag corresponds to (Springer et al., 2008, p. 18). In the case of still image indexing, the existing problems for text indexing are even multiplied (Matusiak, 2006, p. 294) due to the semantic richness and ambiguity inherent to pictorial representations.

Nevertheless, it may be worthwhile to look for ways to surpass these disadvantages, since the application of social tagging may engage audiences and augment awareness of heritage collections (Springer et al., 2008), create different access points (Lu et al., 2010, p. 764; Thøgersen, 2013) that help increasing indexer-searcher consistency, and may complement automatic annotations (Freiburg et al., 2011). One initiative to improve tag quality is *nichesourcing*, a form of human computation which takes advantage of social tagging but involves experts, as opposed to *crowdsourcing*, in which taggers are the general public with no specific knowledge of a given domain (Boer et al., 2012).

Expert and novice generated tags

Social tagging has been defined as a way of organizing information by novices as opposed to the way indexing experts do (Peters, 2009, p. 1). One of the key factors in the success of social tagging in engaging different types of users is the reduction of intermediary steps followed in traditional indexing practices, saving the user from the need for first thinking on a concept and then representing it through the correct term from a controlled vocabulary (Halpin, Robu, & Shepherd, 2007). Different studies compare socially generated tags by non-expert users with the metadata created by indexing experts (Gligorov et al., 2011; Lu et al., 2010; Matusiak, 2006; Springer et al., 2008; Thøgersen, 2013; Trant, 2009b).

In our study, we focus on the relation between the types of generated tags and the participants' knowledge of the domain. Tsai, Hwang, & Tang (2011) looked at whether experts can provide a more consistent and representative set of tags for academic and scientific documents than novices, in the context of nanomaterial technology. They concluded that tags chosen by experts yielded better similarity and relevance values in all analyses and that these tags reflected better understanding of the content. Another study, in the radiological domain by Wang, Ni, Hua, & Chua (2012) explored how novices, intermediates and experts would describe medical images, finding that experts used more high-level image attributes that required high reasoning or diagnostic knowledge than novices, and that novices are more likely to describe basic objects that do not require much radiological knowledge. But Ådland & Lykke (2012) also found that tags can improve the interaction and communication between layman users and domain experts in a domain-specific setting (health information), helping to bridge between scientific terminology (and viewpoints) and everyday problems reflected in non-expert users' vocabulary.

Kang & Fu (2010) take this distinction a level further, by observing not only the tags or the tagging process of these two groups, but also the exploratory information search behavior of experts and novices using a social tagging system, in comparison to a general search engine. They found that expert-created tags could support the understanding of a topic by novices and increase their exploratory search. Closer to our research approach is a small-scale study by Darvish & Chin (2010), comparing film experts and novice tags in a video labeling setting, finding that expert tags were judged to be more relevant by both experts and non-experts and that non-expert viewers also created significantly better tags than the uploaders of the videos.

Guided tagging

For achieving consistency and quality in the tags, different studies explore mechanisms for guiding users in the tagging process. For instance, Smith (2007, p. 128) identified three categories of tag "suggestion systems": previously used tags (suggestions or recommendations based on a user's prior tags), popular tags (based on frequently used tags by others), and recommended tags, suggested by tagging systems based on their own criteria. Faceted tagging is another way of guiding the tagging process, by indicating different aspects of a resource that could be tagged (Smith, 2007, p. 76). For instance, Bar-Ilan, Shoham, Idan, Miller, & Shachak (2008, p. 941) found that structured tagging, which guides the user by presenting "fields" (such as "event, symbol, personality, date, place"), usually resulted in more detailed descriptions. In a practical application, the "Your Paintings" tagging project (4) applies this in practice: it guides users when tagging different aspects of a picture, such as things, people, places, events, subjects, and types. Sen et al. (2006) showed in an experiment on vocabulary formation in the Movie-Lens system how different design choices affect the nature/types of tags used, their distributions and the convergence within a group.

In sum, as Good, Tennis, & Wilkinson (2009, p. 14) point out, investigation on methods for guiding user contributions in particular directions is an important area of tagging behavior research. In the experiment we describe here, a group of randomly selected taggers received guidance in the form of instructional text informing about the types of tags which should be used.

Tag categories and models for image description

Although active research on tag categories exists (Peters, 2009, p. 196), to our knowledge, there are no studies about the different types of user-generated tags in a time-based fashion within the audiovisual domain. In our study, with the aim of creating an instructional guide on tag types for film content, and of observing semantic categories and types of tags used by expert and novice groups, we selected four types of tags by combining different models for fixed image analysis found in the literature. In its review, the Panofsky/Shatford matrix was found to be a widespread model for describing image content (Westman, 2009, p. 64). Panofsky (1977) addressed the levels of meaning in artistic images, defining three properties: pre-iconographical, iconographical and iconological. Layne (1986) followed with an extension of Panofsky's theory, adding four more facets (who, what, where, when).

Further, Hollink, Schreiber, Wielinga, & Worring (2004) adapted, extended and applied some of their preceding models for creating a framework that was used for classifying visual resources related queries and annotations. The framework distinguishes three viewpoints on images: the non-visual metadata level, the perceptual level, and the conceptual level.

More recently, Tirilly et al. (2012) proposed a model of image description based on characteristics obtained from experimental data in a study about the features of image similarity. According to them, their model provides a basis to define the image features that image retrieval systems should implement (p. 170). The features in their model refer to the image properties (e.g., type/technique, focus, point of view, lighting, contrast, file quality), to the scene's semantic and physical properties (e.g., place, time, color, composition), and to the objects' semantic and physical properties (e.g., nature, emotion, color, texture).

Golbeck, Koepfler, & Emmerling (2011) applied the Panofsky/Shatford model to the analyses of the social tagging behavior of image content. They tried to discover the relationship between tagging behavior and the features of the images which were tagged. They found that users' past experience with an image as well as the type of image being tagged creates significant differences in the number, order, and type of tags (p. 1750).

Even though the models mentioned above refer mainly to still-image analysis, they have been used to analyze moving images as well. Hollink (2006) used her framework for classifying visual resources (Hollink, 2006; Hollink et al., 2004) in three different contexts, one of them being broadcast news for a content-based image retrieval system. The results showed that the specific level was more important in the news domain than in the other domains (p. 121). In turn, Gligorov et al. (2011) used Hollink's and Panofsky/Shatford models in the analysis of *Waisda?* tags for television programs of a broad and entertaining nature.

In a study of key-frame extraction, Kim & Kim (2010) reviewed six representative models for still image analysis, concluding that people interact with images at three levels: primitive features (e.g., color and shape); derived attributes (e.g., specific objects), and semantic abstract attributes (e.g., the symbolic value) (Greisdorf & O'Connor, 2002)", which resemble the three panofskian levels.

In general, we found a lack of research about how these models for still image analysis can be applied or adapted to moving images, and observed a gap in the literature in identifying the formal and content attributes of time-based descriptions that are meaningful to expert and novice users.

Experimental design

Research questions

RQ1. How do film experts tag films compared to the general public? Do film experts, as opposed to novices, reflect their domain specific knowledge when tagging film content?

Tags are a spontaneous way to associate words with digital content, which reflect the users' personal understanding of a topic or their intentions with the digital resources (Tsai et al., 2011). For that reason, we might hypothesize that domain experts would use their domain-specific terminologies when tagging. We thus study the types of film experts' tags and compare the differences between film experts and novices when tagging film content in a realistic *crowdsourcing* environment. We analyze, among other things, the distribution of their respective contributed tags through different semantic levels.

RQ2. Can we influence the type of time-based tags that users enter with specific instructions?

One of the problematic issues of indexing/tagging audiovisual content is that there are many levels or dimensions of meaning involved. To address this question, we investigate if experts and novices enter more specific tags when they receive instructions of using different semantic categories that may apply to film content.

Test procedure

To address our research questions, we designed a 2 × 2 between-subject study for which two groups of participants were selected: film experts and domain novices. In turn, these groups were divided into two sub-groups: one having instructions (guidance in which types of tags they could use), and the other one having only general indications on how to play the game, but no instructions on the types of tags to enter. .

All participants were asked to play a game with each of the five videos. Since we were interested in the types of tags, they were allowed to use their mother tongue when tagging if it was English, Dutch or Spanish, with the aim of favoring their spontaneity. The participants were asked to fill in a questionnaire after completing the five games.

Selection of participants

In total 36 persons participated in this study: 18 film experts and 18 domain novices, 9 out of the 18 in each group received instructions and 9 did not. The participants were selected in two different ways:

- **The film experts**. We considered people involved with film content at a professional or academic level and linked to film-related institutions. Our participants were contacted in film and television archives, universities, a government institution, and at a national library's film archive. They were based in The Netherlands, Norway, United States, Spain, and Colombia. In total, 45 invitations were sent, and 18 experts completed the full experiment (response rate: 40%). This group included participants who were film historians (scholars), cataloguers or archivists (curators), filmmakers, film/video technicians and film programming staff. All of them had an academic background in and/or formal education related to cinema. The age of the experts was between 30 and 39 (n=12), 50 and 59 (n=3), 20 and 29 (n=2), and 40 and 49 (n=1). Half of the participants had working experience with film materials and content of 10 years or more (n=9); between 7 and 9 years (n=6), 4 to 7 years (n=2), and one was a junior researcher (less than 3 years of working/research experience). There were twelve females and six males.
- **Film novices.** As non-domain experts, we considered people without a professional or academic relation to film content, and people not familiar with terminologies related to film. They were recruited by using an informal call for participation on one of the author's Facebook pages, indicating that not being a film expert or enthusiast was the only requirement. In total, we got 26 positive replies. From those, 18

completed the full experiment.

The novice group consisted of professionals with high-level education, mainly with a Library and Information Science background. This indexing expertise factor was not intentionally sought in the study, but since we were interested in domain specific knowledge we did not consider it a problem, rather we saw it as an advantage, since it helped us have a higher number of participants in all groups with knowledge and experience with tags and keywords. Regarding their ages, most novices were between 30 and 39 (n=9), the others were between 20 and 29 (n=5), 40 and 49 (n=2), and 50 and 59 (n=2). All novices defined themselves as such, that is, their domain-specific knowledge or distinct concern about films was null, and their interest in them was not explicitly reported to go beyond occasional movie-going activities. There were fourteen females and four males.

Prototype application

We used the *Waisda?* system (5) for the experiment setup. This is available as free and open source software at the GitHub repository (6). Figure 1 shows a screenshot of the tagging interface where it is possible to see how tags are entered while the video plays, being attached to a specific time point in the video. Users get points by entering tags, and a higher score when the tags match with the tags entered by other participants. A detailed explanation of the software, game rules and interface is described by Hildebrand et al. (2013).

"Insert Figure 1 here"

For the experiment, the functionality of *Waisda?* was modified in two ways. Firstly, we neutralized the effect of the game scores on the tagging behavior: points were not given when the tags entered by one participant matched with tags entered by other participants. This was done to prevent participants from entering the types of tags that will maximize their score. For example, if a player observes that by entering 'woman', (s)he is rewarded with points, then (s)he would be encouraged to enter other tags of that type, such as 'man', 'dog', etc. This is what Fu, Kannampallil, Kang, & He (2010) called "semantic imitation", where "users who can see tags created by others tend to create tags that are semantically similar to these existing tags". Semantic imitation is an important characteristic of tagging games, but for the purpose of our experiment it had to be neutralized. As a solution, we decided to retain the scoring mechanism of the game, but to control the tags that are rewarded with points, in order to guarantee a fair distribution over the different tag category types (see "user instructions" section). Therefore, we introduced a single non-real player (a bot) that all the participants competed against, exerting the same influence on all subjects. The players were rewarded with points for matching with tags of the bot, but were unaware that they were not competing with other players. For each of the five film clips, we created a set of tags for the bot that covered each of the five tag categories included in the instructions. In this way, participants were rewarded for entering matching tags in the different categories, and not only for factual tags, which was the most common tag type in previous *Waisda?* experiments.

As a second modification, we disabled the display of tags entered by other players in the *Waisda?* game and on the *Waisda?* homepage, to neutralize all tag suggestions other than the instructions of the experiment.

Selection of film clips

We uploaded five clips from the European Film Gateway (EFG) (7) into the system. The EFG is a portal that provides access to the digitized collections of around sixteen European film archives and cinémathèques. We made a purposive sampling by selecting five clips according to the following criteria:

- They should be from films with no dialogs, with the aim of avoiding script transcription as much as possible;
- They should be short (no longer than five minutes), as previous *Waisda?* studies had indicated that the players prefer playing games with short clips.

Except for a Swiss short film, our final selection included movies from renowned Danish and German film classics or directors; we also assumed that if these movies were presented at the EFG their value was previously assessed. The five selected film clips were (clip duration is between brackets, and a link to the EFG record is referenced): "Den flyvende cirkus" (Alfred Lind, Denmark, 1912; [02:02]) (8), "Die Gezeichneten" (Carl Th. Dreyer, Germany, 1922; [00:37]) (9), "L'aiguille" (William Piasio, Switzerland, 1961, [05:55]) (10), "Metropolis" (Fritz Lang, Germany, 1926, [01:30]) (11), and "Vampyr" (Carl Th. Dreyer, Germany/France, 1932, [01:36]) (12).

Participants' instructions

All participants received a common set of instructions by email, indicating how to play *Waisda?* (also available on the Waisda/EFG homepage). Participants that were part of the "instruction group" received another set of instructions, with details on the types of tags they could use (see "classification No.1" in the "Data analysis procedures" section). We created a simple "instructional model" based on some features of the models described in the section "tag categories and models for image description". The following were the resulting instructions that we provided to the participants:

> "Tags consisting of one or two words are more likely to match than longer phrases. Tags may be about the following aspects (please try to cover as many as you can during the game):
>
> - **Facts**. What you see or hear in the scene, such as objects, persons, places and actions (e.g. woman, sofa, London, R2D2, murder).
> - **Cinematography**. Stylistic features, such as form, style, framing, camera movement, lightning key, type of shot, camera angle (e.g. backlighting, wide-angle, caligarism).
> - **Explanations**. Symbolic interpretation of the meaning or theme (e.g. psychotic rage, oppression, dehumanization).
> - **Emotions**. The emotions, thoughts or intentions of the characters (e.g. bored, despair) or your own emotions (e.g., fascinating).
> - **Other.** You can use other types of tags that are not described here".

We did not intend to create a "new" model or set of categories in this text, but rather interpreted and summarized some of the important features pointed in the existing models for image analysis related to film content. For instance, the "Facts" category, is inspired by Panofsky-Shatford's 'pre-iconography/ generic 'of' and Iconography / specific 'of'', and in Baca's (2002) 'Ofness' categories. Our "Emotions" concept coincides with Panofsky's (1955) 'Pre-iconographic (expressional) category' and other models which consider emotional abstraction (Eakins, Briggs, & Burford, 2004). Our "Explanation" type was derived from Panofsky's (1977) 'Iconology' category and Ingwersen's (1992) 'aboutness', and our "Cinematography" type from Hollink, Schreiber, Wielinga, & Worring's (2004) 'perceptual' category and from one of the key books on cinematography (Bordwell & Thompson, 2003).

Questionnaire

The participants were asked to fill in a questionnaire after completion of the test (13). The questionnaire consisted of 22 questions, divided into three sections: demographic information and expertise level; previous experience with indexing, tagging and labeling games; and the participant's experience with the game and experiment. In this last set of questions, participants were asked to rate their level of difficulty in coming up with tags, the influence that scoring in the game had on their motivation, the usefulness of the instructions, and their perception of the value of their tags for future use. The participants were also asked to select the types of tags (factual, emotional, etc.) they used, according to their judgment. There were also open questions in which participants could write their comments about these different aspects.

Data analysis procedures

We omitted tag stemming procedures since we are mainly interested in the type of tags that were entered, and not in the matching tags or tags morphology. All tags entered in Spanish and Dutch were manually translated into English, and misspellings were corrected, only with the aim of facilitating the tag category analysis (14).

In the quantitative analysis of the tags, we consider the number of tags that were entered. In this experiment, we do not include precise quantitative results of matching tags, due to the presence of tags in different languages. In the semantic analysis of the tags, in order to analyze their types, we manually classified them according to four different tag classifications (Classification No.1 corresponds to the instructions given to the participants, while Classifications No.2 to 4 were used for complementing the analysis but were not provided to the participants. In these last three classifications, we followed the same approach as in Gligorov et al. (2011)):

- *Classification No.1***.** “Instructional model” (**Facts**, **Emotions**, **Explanations**, **Cinematography**, **Other**). For the criteria to classify a tag in these categories, we used the examples and descriptions given to the participants, and we added some criteria for classifying the data.
- *Classification No.2*. “Hollink’s model”, also as used in Gligorov et al. (2011) includes the **Non-visual level** (descriptions that are meant to describe the context of the video but not its content); the **Perceptual level** (tags that are derived from low-level audio and visual features of the video); and the **Conceptual level** (tags that describe the content of the image, giving information about the semantic content of the image). We only use this classification to filter out the conceptual tags.
- *Classification No.3***.** “Panofsky-Shatford model (specific, abstract, general),” as used in Gligorov et al. (2011). At this level, tags that were classified as conceptual are classified according to their specificity level into specific, abstract or general. **Specific** (iconography) tags possess the property of uniqueness, for example, the name of a person or place. **Abstract** (iconology) tags are those which level of subjectivity allows for differences in opinion, (e.g., ‘crazy woman’). Also tags expressing relationships

(e.g., 'friends'), or tags related to occupations (e.g., 'artist'). The last category is **General** (pre-iconography), which can be derived from the visual properties of the image or sequence alone. Tags classified as General do not have to be correct ('dog barking', and 'duck quacking' were used in the same time frame, this low level of subjectivity is not enough to consider the tag Abstract).

- *Classification No.4***.** "Panofsky-Shatford model (who, what, where, when)." We used the concepts from Shatford (1986): "**Who**" refers to the concrete objects and beings, animated or inanimate; or individually named persons, animals, things; or to kinds of persons, animals, things; or to mythical beings, abstractions manifested or symbolized by objects or beings. "**Where**" to a location and "**When**" to time. "**What**" refers to an event in the video: "what are the objects and beings doing? (action, events, emotions)", explains Shatford.

The tags were manually classified by one of the authors. A sample of the tags was classified by a second person for assuring the consistency of the classification criteria. We used a quota sample by randomly selecting tags created by each of the four subgroups for each video. The Cohen's kappa (k)[2] was used as a measure of agreement between both annotators. The results were reasonable for three of the classifications (0.67 for 1 and 2, and 0.62 for classification 3). The agreement was low (0.32) for classification 4. However, more in-depth analysis showed that this was due to a different interpretation of the Panofsky-Shatford's model in relation to the "Who" and "What" categories, which are explained differently in the original Shatford (1986) model, and in Gligorov et al (2011, p. 150). This does not reflect a disagreement in the tags classification but a different interpretation of the model. Since it was applied systematically in the classification by each annotator of a small proportion of tags, we concluded that the categorization was consistent and not arbitrary and that we could use it for analyzing our results.

After tag classification procedures, we manually clustered synonyms and singular/plural forms to look at the most frequent types of tags from a semantic perspective (the tags obtained from these clusters were used in **Error! Reference source not found.** and 4). Finally, we analyzed the answers to the questionnaire: 1) to help interpret the results of the quantitative and semantic analysis, and 2) to discover the participants' perceptions of tagging behavior.

Findings and discussion

Next, we present the findings from the examination of the tags and the analysis of the questionnaire answers.

Number of tags

The 36 participants contributed a total of 2,943 distinct tag entries for the five videos. 2,404 were in English, 262 in Spanish, and 276 in Dutch. From the 2,404 English tags, 1,137 were unique. Error! Reference source not found. shows the means and standard deviation of the tags entered by each group. The high standard deviation among the participants in group D (58.1) was due to the presence of one "super-tagger" (as called by Trant (2009b). However, we did not detect any outliers (using the outlier labeling rule with a value of 2.2 as the multiplier).

"Insert Table 1 here"

A Kolmogorov-Smirnov test showed that tags per group and video were not normally distributed. We therefore chose to conduct a Kruskal-Wallis test (a non-parametric test for independent samples and three or more groups) to examine the relationship between number of tags, expertise and instructions among all groups, as well as a Mann–Whitney *U* test for testing differences between pairs of groups.

The results showed that, in most cases, there is no effect of expertise and/or instructions in the number of tags entered by the different groups (*p*> 0.05). One exception appears in the evaluation at the individual video level, for which there was a significant difference for the clip of "Metropolis": i) in the number of tags entered between all groups (*p*= 0.013); ii) between the groups A and C (*p*= 0.019); and iii) between the groups B and D (*p*= 0.024). We will comment on this later.

Types of tags

To observe the types of tags among the different groups, we used "Classification No.1". As we can see in Error! Reference source not found., the distribution of the types of tags among the different groups shows that all of them predominantly entered "Factual" tags. To illustrate which tags belong to each category, Error! Reference source not found. includes the three most frequent tags per group**.**

"Insert Figure 2 here"

"Insert Table 2 here"

"Factual" tags correspond to objects or actions that are depicted in the scenes. These "ofness" words (as defined by Baca (2002); Peters (2009); Layne (1986)) correspond to what Panofsky calls the "pre-iconographical" level of meaning: the description of "primary or natural subject matter", which is apprehended by identifying *pure forms* (Panofsky, 1977, p. 5). Even though object identification is not a simple process (from the semiotic point of view), it is assumed here that these descriptions do not require film domain specific knowledge.

To examine closer what happened in the other four tag categories, and for observing the effect of expertise and instructions in the distribution of the types of tags, we performed a Kruskal-Wallis test again, and a Mann–Whitney *U* test for testing differences between pairs of groups. Error! Reference source not found. shows the cases with a statistically significant difference (p < 0.05).

"Insert Table 3 here"

In Error! Reference source not found. we observe that there is a significant difference in the use of tags of the type "Emotion" between all groups, and by almost all the analyzed pairs of groups. This result was not expected. The group of experts with no instructions (A) had significantly fewer tags of the type **"Emotions"** than the respective novices group (C) (5.77% vs. 11%, p=0.003). In turn, the groups with instructions (B and D) entered more tags of this type than their counterpart with no instructions (A and C) (10.48% vs 5.77% p=0.024 for the experts groups, and 15% vs 11%, p=0.031 for the novices groups).

This difference may be caused by the level of awareness that the instructed groups gained on this type of tag. "Emotional" tags correspond to feelings expressed by the characters in the scenes as detected by the taggers (e.g. 'angry'), or to feelings experienced by the tagger her/himself (e.g. 'creepy'). The last type coincides with what Zollers (2007) identified as "opinion tags". Normally, the use of emotional attributes is not prescribed by traditional cataloguing or indexing guidelines. However, there is growing interest in the structured identification of emotional aspects from various art forms for different purposes (e.g., movie recommendation). Affective tagging could be used as part of user engagement activities (e.g., as in the "Emolab project" (15)), and/or for retrieval based on non-factual information during footage finding or research. For instance, Inskip, MacFarlane, & Rafferty (2008) describe the process of searching for accompanying music to film scenes, which involves highly subjective affective meanings, where motional tags could be useful. In turn, there is active research in the psychology domain (Bálint & Kovács, 2012) and in film studies (16) about the emotional involvement of the film viewer, which require or benefit from this type of tagging.

Also, in Error! Reference source not found. we can observe a predictable result in relation to “Cinematographic” tags: a significant difference (*p*= 0.024) in the number of tags entered by group B in relation to group D (7.76% vs. 3.54% of each group’s total tags, as it can be seen from the proportions in Error! Reference source not found.). **“Cinematographic”** tags correspond to domain-specific terms, such as photographic aspects of the shots or framing, camera movements or editing characteristics. In relation to RQ1, on whether experts’ tags reflect their specific knowledge, we expected that the lack of domain-related knowledge made it difficult for novices to describe their cinematographic aspects and that this type of tags would be more used by experts. Unexpectedly, novices also used this type of tags, but in a more general fashion than experts did (for instance, as shown in Error! Reference source not found.**,** by using tags such as ‘black and white’, or ‘silent film’). In relation to our first question, about how experts and novices’ tags differ, Error! Reference source not found. confirms an important distinction, which is the experts’ variety of domain-specific terms in relation to cinematographic language. These terms are located in the long-tail portion of the expert tags’ distribution and are thus not quantitatively significant, but semantically rich from a qualitative perspective.

“Insert Table 4 here”

We explored the semantic overlap of this tags’ sub-set with The International Federation of Film Archives (FIAF) thesaurus (17), looking for similarity (syntactic and semantic) between the sample of tags in Error! Reference source not found. and the thesaurus descriptors. From the 77 Cinematography tags, only 10% (n=8) had an exact equivalent (syntactic and semantic); 32% (n=25) had some equivalent in the thesaurus (e.g. for the tag ‘silent film’ the correponding term would be “history of cinema. silent period”). None of the tags indicating shot type were found in the thesaurus, where the broader terms “Camera angles” or “Cinematography” cover all the spectrum. Other indexing alternatives different to thesaurus-based indexing are extensively investigated by Turner (e.g., 2009).

We assume that there are richer semantic connections within the tags themselves, and not only in relation to external vocabularies that do not have a time-based focus. In this sense, a relevant topic for future work is mining the semantic associations between tags and tag provenance in relation to the time dimension. For example, within a 10-second span, we can have a combination of expert and novice tags such as ‘abandoned’, ‘house’, ‘panning’. If the tag ‘panning’ was added by a film expert, this could eventually indicate that there is a pan shot of an abandoned house in that time frame.

From Error! Reference source not found., there does not seem to be any significant difference between the groups in the use of the tags of the type **“Explanatory”**. These tags range from the simple registry of objects and actions, to the higher level of abstract ideas, symbolic interpretations or interconnections (for instance, finding a relation with an art or literary movement, as in the tag ‘expressionism’). These tags require from the tagger more effort in using her/his background knowledge, whether film related or not. In our test, both film experts and novices provided this type of tags to a low extent.

The **“Other”** category also lacks a significant difference. These tags mostly correspond to what in Classification No.2 is categorized as “Non-visual” level. It covers descriptive metadata such as the date (e.g. ‘1912’, ‘1932’), location or country of origin (‘french movie’ ’), creator (e.g. ‘Dreyer’), title (‘metropolis’), or historical-contextual aspects (e.g. ‘early cinema’).

Following the procedure used in Gligorov et al. (2011), we used Classification No.2 to filter out only the “conceptual” tags for the subsequent Panofsky-Shatford analysis (Classifications No.3 and 4). Tags classified in this category (“conceptual”) corresponded to 86% of the tags’ total (coincidentally this proportion is almost the same one found by Hollink (2006), who concluded in her empirical study about the use of the different categories in her model –our Classification No.2- that the conceptual levels were used most (87%)). Error! Reference source not found. shows the proportions of “conceptual” tags for the most frequent Panofsky/Shatford categories.

“Insert Table 5 here”

In relation to RQ1, the figures in Error! Reference source not found. confirm our previous finding of the lack of substantial dissimilarities in the most common semantic types of tags by both groups. In this case, both experts and novices used more tags of the type “General/Who”, with no significant statistical difference between groups. This category corresponds mostly to factual tags and more specifically, to descriptions of objects in the scenes. This result agrees with Thøgersen (2013) who found in his study about still image tagging by general users that most tags were of the type “Artifact/objects”. After this category, tags in the “General/What” category predominate; these are descriptions of what happens in the scenes at a general level (e.g. ‘bell ringing’).

“Abstract/What” tags were the third more used type by both experts and novices, which corresponds to descriptions of events or actions in the scenes at an abstract level (e.g. ‘calamity’). In this category there was a statistically significant difference between groups A and C: Error! Reference source not found. shows that non-instructed novices (group C) tended to use more “abstract/what” tags than non-instructed experts (group A) (26.37% vs 15.09% respectively; *p=0.031* after a Mann–Whitney *U* test). These tags coincide with explanatory and emotional tags, which are of a more abstract nature.

“Insert Table 6 here”

In relation to RQ2, about the effect of instructions in the tags’ selection, we found that instructed experts (group B) tended to use more abstract terms than their counterpart group without instructions (group A) (*p= 0.040,* from a U Mann-Whitney Test for groups A and B in the abstract category using Classification No.3*)*. This difference was due to the increased use of “Abstract/Who” tags by the instructed expert group (B) in relation to the non-instructed expert group (A) (*p=0.031,* using values from Error! Reference source not found.*).* The experts’ preference for general tags over abstract tags (**Error! Reference source not found.**) shows similarities with conclusions reached by Thom-Santelli, Cosley, & Gay (2010). In their study about the differences between experts and novices in a collaborative environment, they found that experts have a preference for objective tags. The preference for this type of tags in a video labeling game also agrees with Gligorov et al. (2011), who found that most conceptual tags were general (74%). In our test, percentages of “abstract” tags were higher (31% of the total conceptual tags) than in Gligorov’s study (7% of the total conceptual tags). This difference may be caused both by the type of content (film in our study vs. television in their study) and/or by the guidelines given to the taggers, which included “Emotions” in the possibilities.

Perception of the value of instructions

Participants in the guided groups (B and D) were positive about the value of instructions in helping them to come up with tags. Error! Reference source not found. shows that when asked about the value of the given instructions (q18) (18), the median from groups B and D is higher than for the non-instructed groups (A and C).

A higher value of instructions was perceived among the novices group (D).

“Insert Table 7 here”

A number of non-instructed experts and novices (n=5) suggested that the categories that we used in the questionnaire to ask them rank the types of tags they used (Facts, Emotions, etc.) (q16) could have been used in the instructional text as guidance for which types to use. These reactions indicate that instructions about types of tags are necessary for time-based tagging. Participants described in the open answers to the questionnaire several issues which can be summarized in these points: (a) taggers need to know which aspects or dimensions they should focus on during tagging; presenting several types of tags in the instructions may help, but the participant needs only one to keep the focus; (b) participants should have previous knowledge about the movies and clips (e.g., contextual or historical information, and information about the clip itself), (c) the future retrieval purpose of the tagging activity should be stated; and (d) term suggestions may help the tagger.

The role of professional experience with indexing, tagging and labeling games

Lee, Goh, Razikin, & Chua (2009) showed that “the familiarity of users with the concept of tagging, the functionality of tagging systems, and the use of web catalogs has a great effect on the user’s tagging behavior” (p.184). To observe these issues, we asked the participants to rate their level of professional experience with indexing/cataloguing, their familiarity with creating tags, words or keywords for online content (for example: labeling images in Flickr, or videos in Youtube, or bookmarks in Delicious); about their familiarity level with video search through keywords or tags, and their knowledge and experience with video labeling games.

However, we did not find a statistically positive correlation between the number of tags entered by the participants and each one of these different aspects (using the Spearman’s Rho two-tailed test). This may be attributed to the quite homogenous “indexing” expertise of our participants regardless of their domain expertise. This leads us to be cautious about concluding that our study contradicts results from Lee et al. (2009), but rather that testing mechanisms for tagging familiarity should be refined in future tests.

The influence of content, and familiarity with the content

As expected, the domain expert participants reported familiarity with some of the video clips: (1) with “Metropolis” (n=15 high familiarity, and n=3 medium familiarity), “Vampyr” (n=7 high familiarity, and n=4 medium familiarity); and “Den flyvende circus” (n=1 high familiarity, and n=3 medium familiarity). There was a positive statistical correlation between the most familiar clip for all participants (“Metropolis”) and its total number of tags (*r=0.442; p=0.007 from a* Spearman’s Rho two-tailed test for testing correlation between familiarity with each film and its corresponding number of tags for this clip), which indicates that a higher level of familiarity resulted in more tags. There is also a negative correlation between familiarity with this film and the use of emotional tags (*r=-0.461; p=0.005)*, which indicates that the more familiar the tagger was with this film, the less likely was to use emotional tags. This corresponds to our previous findings of a marginal significant difference in the number of tags at the video level for the clip of “Metropolis”. In this case, the experts’ groups entered more tags than the novices’ groups, and those tags were significantly less of the type “Emotions”. Instead, “Explanations” and the “Other” tags’ categories were more used, which reflect experts’ knowledge about the metadata attributes and interpretations of this movie (e.g., ‘dystopia’, ‘Fritz Lang’).

From the participants’ answers, it was also observed that familiarity with the content plays an important role in motivating the tagger. It also allows the participant to concentrate on tagging, and not on getting acquainted with a movie that is new for her/him. As one expert states: *“there is always the difference between knowing a film and seeing it for the first time. The first time [you have] reactions on what you see, the second time is more intentional” (Participant group B).*

Game effect, scoring and tagging motivations

A common feeling among the participants from all groups was time pressure. They found that the short duration of the clips, or the impossibility to replay them, added stress to think of, or limited them to entering more tags, both during the video (because they were watching it and not entering tags) or at the end of the clip (tags for the last frames). One expert commented that this was not “a professional way of working” *(Participant group A)*.

“Insert Table 8 here”

From Error! Reference source not found., we can conclude that it seemed to be easier for the experts groups (A+B) to come up with tags than for the novices. Among the instructed experts group (B) there were participants dissatisfied for not being able to enter all tags that occurred to them. They explained that the lack of familiarity and short duration challenged them in this respect. Participants from different groups pointed to different negative issues related to the game influence. These include (a) “multitasking” (i.e. watching the video, thinking of tags, typing it in); typing skills (having to look at the keyboard); (b) the impossibility to synthesize in a single word or in a couple of words the concepts they had about the fragments, and/or to recall the technical terms referring to shot types and editing; (c) language issues and spelling.

The reaction to scoring and gaming elements (q15) are very personal, and we cannot conclude any relation to domain expertise. Some experts made positive comments about the game itself and found it fun. Both among the experts and novices groups there were few participants concerned for having few matching tags. Not surprisingly, we found a positive correlation between scoring motivation and number of tags (*r=0.406, p=0.014* after a Spearman’s Rho two-tailed test). A drawback of this correlation, also identified by Thøgersen (2013), is that since the game is set up to reward players based on matching tags, this encourages most players to tag what is in the picture, rather than thinking about other possibilities.

Finally, as in other tagging activities, there should be a quality control and feedback mechanism that allows the participant to check the value of her/his tags. One novice said: *“It was very easy to write a tag when it came up in mind. The only difficulty was in deciding if it was a “correct” tag, i.e. if the word made sense or it was just an instinctive reaction to what I was seeing” (Participant group C)*.

We can conclude that clear guidance and objectives in the tagging activity, encouraging participants to use their specific domain knowledge, and a flexible tagging setting (not necessarily competitive), may increase the motivation in the tagging activity beyond scoring mechanisms. Future work should focus on investigating which rewarding mechanisms work better for experts. One direction is suggested in the study by Thom-Santelli et al. (2010), who points to innate experts’ feelings of territoriality and “curation”, which means that experts can have higher levels of participation due to ownership feelings in cooperative work that involves targets of their concern (e.g. museum objects).

Perception of the utility of selected tags

Error! Reference source not found. indicates that novices were more positive about the possible use of their tags for future retrieval of the videos than experts, who were mostly uncertain.

"Insert Table 9 here"

Indeed, domain experts cast doubt on the tags' semantic value. They consider them very general and only related to describing what they saw in the images, without taking into account any context. For these experts, this does not correspond to describing the actual content of the film. For instance, one expert stated: *"My tags were very factual, about what you see in the image. If you want footage of a train, then you will find L'aiguille. If you are looking for a silent expressionist horror film, you will not find Vampyr with my tags" (Participant group A)*. One more expert confirms the utility of her/his tags, but, as s(he) says: *"only for such purposes as stock video footage, but not for meeting thematic or content driven curatorial or research needs" (Participant group A).*

This shows the need for more research in understanding the use of time-based annotations for research purposes, beyond footage finding. From the novices perspective there are other concerns, one novice commented: *"I guess moviegoers tend to select films based on the genre as well as actors/actresses and maybe directors involved with the film. I am wondering how social tagging plays a part in helping us decide which films to watch" (Participant group D).* Current practice is showing interesting directions in involving humans in creating keywords for movie recommendation for entertainment, such as the Netflix case described by Madrigal (2014). These practices have roots in cultural heritage curation, and film archives can benefit from them for dissemination purposes.

Limitations

The data collection took place in a game setting, which may be a very specific type of tagging scenario. However, even though this study did not include a comparison between the differences with non-game contexts, most of the findings were in line with conclusions found in other experiments based on other data collection methods.

In relation to homogeneity in the experts and novices groups, we did not include in our procedures a detection and/or operationalization of expertise by testing the actual knowledge of the participants (as it is done for instance in Kang & Fu, 2010). Additionally, we omitted any form of control in the participants who got the instructions to know if they read them in detail; at least one participant admitted to having skipped a careful reading.

About the labeling setting, we chose to let participants play against a bot, instead of the default setting: against each other. Influence in tag selection by the participants is, in both cases unavoidable and difficult to judge or measure.

We find the procedure of allowing taggers to use their mother tongue valid for our research purposes, but multilingualism is far from being a trivial issue, and a research area on its own that we did not touch in depth it in our study.

Finally, this was a small-scale experiment that counted with the participation of the minimum number of film experts and novices (45 cases per group: 5 videos x 9 participants). A higher number of participants would be needed to validate the findings quantitatively.

Conclusions and future work

Coming back to RQ1, about how domain experts tag film content in comparison to novices, we observe that experts tag in a similar fashion as novices when participating in a tagging game. In general they enter the same number of tags, and they mostly use “Factual” tags. However, in the experts’ less-frequent tags, we found more domain-specific terminology than in the novices groups.

The use of the most common type of tags (“Facts”) among the two groups, agrees with other studies on image subject categorization (Klavans, LaPlante, & Golbeck, 2013), with other game related experiments (Thøgersen, 2013), and with the tag analysis of the first *Waisda?* projects for TV broadcasts. These tags describe the content at a general level (Gligorov et al., 2011). Perhaps, as Halpin et al. (2007) indicate, tagging requires less cognitive effort, which would explain why experts tagging behavior was similar to the one of novices, at least under the conditions given the chosen labeling setting. And yet we think that a clearer explanation for the groups’ similarity is the competitive nature of the game. We confirm this is not the best scenario to tap into the domain-specific knowledge of experts (as it was somehow expected, and also pointed out by the experts themselves in their comments). However, the same game has proved to be useful for getting a relatively high number of relatively high quality time-stamped tags from general users (as Ahn & Dabbish, 2008; Gligorov et al., 2013 found out). This poses the issue of how to join the advantages of a great number of common tags (which can improve indexing consistency, assumed to indicate quality (Good et al., 2009, p. 6)) with less frequent expert tags, assumed to be more relevant for specialized contexts (Tsai et al., 2011). In this regard, we confirm the need for extracting tag provenance information, which can add to the quality measures of the tags. This follows the tendency to mining not only the relationships between tags and documents, but the link between users, tags, and documents (as suggested by Good et al., 2009). In addition to this, one aspect that was not possible to cover in this study, but which needs future exploration is the analysis of the influence of film genre in the types of tags.

In what concerns RQ2, about the influence of guidelines in the selection of types of tags, we conclude that novices can provide tags in different types of categories. However, as expected, the level of detail of the individual tags in the most domain-related category (Cinematography) is limited. Additionally, from the questionnaire we know that most participants preferred to have a clear description of the type of tags they were expected to enter. In the case of moving images, where several dimensions co-occur, instructions should help participant focus on specific content or stylistic aspects and allow complementarity of novice and expert tags for the same video. For instance, one of the usage scenarios for online film archives to enrich and give access to their online digital collections could be to ask experts to contribute only “Cinematography” tags. In this way, film experts’ tags could be used for novices in browsing and learning the cinematographic language because expert tags seem to have the potential to augment the exploratory search of information. This holds especially for users who have little knowledge on a topic (as Kang & Fu (2010) found). Novices, on the other hand, should be guided to contribute “Facts” (and eventually emotions or explanations) in their tags, according to expertise in other domains, not necessarily film-related backgrounds. In any case, *nichesourcing* initiatives are about channeling expert knowledge instead of asking experts to do what novices, or eventually content-based retrieval algorithms, could also do.

More studies need to be done to understand the way of motivating and obtaining significant time-based tags or annotations from film experts and novices for research or educational purposes, and not only for footage finding. Current research also points to the fact that tagging implementations should be part of more integrated curatorial and annotation infrastructures, and that isolated tagging support may not be the best solution to obtain expert tags. One example is “The Larm Project” (Skov & Lykke, 2012) in which a national research infrastructure for radio and audio-based research is built through a collaboration between universities and radio archives. This infrastructure would support knowledge dissemination, sharing and interaction between different kinds of humanities researchers, by providing necessary scholarly-based links between texts and images

(Winget, 2009). In fact, in these settings, games are still an option, though not the only one. A requirement is that more varied genres of a higher collaborative nature are investigated, as pointed out in Goh et al. (2011). We are currently exploring if we can use expert descriptions made outside the game setting to improve the tags made by novices inside the game. Also, we are exploring the use of term suggestions from, for example, the IMDB plot keywords (19) database, or from technical film glossaries. Although these techniques are already in use, more theoretical work needs to be done to provide semantic models and classifications schemes specific for moving images.

Acknowledgements

This study was possible thanks to the enthusiastic participation of 36 experts and novices who played the *Waisda?* game and contributed with insightful comments. We also thank Erwin Verbruggen and Hugo Huurdeman for contacts and advice, and our reviewers for the valuable suggestions.

Footnotes

(1). http://www.bbc.co.uk/radio4/science/findlistenlabel/

(2). http://celluloidremix.openbeelden.nl/

(3). http://www.scenemachine.nl/

(4). http://tagger.thepcf.org.uk

(5). http://prestoprime.cs.vu.nl/efg

(6). https://github.com/beeldengeluid/waisda

(7). http://www.europeanfilmgateway.eu/content/about-european-film-gateway

(8). “Den Flyvende Cirkus” at EFG: http://tinyurl.com/p8cutp5.

Film by the Film Fabrikken Danmark production company. Directed by Alfred Lind (1879-1959), whose name is “inextricably linked with a large part of Danish silent film milestones”, according to the Danish National Filmography (http://www.dfi.dk/faktaomfilm/person/da/127597.aspx?id=127597).

(9). “Die Gezeichneten” at EFG: http://tinyurl.com/nhrdpn6

Original title “Elsker hverandre” (Love one another). Directed by Carl Theodor Dreyer, recognized to be Danish cinema’s most important director; (Danish National Filmography, http://www.dfi.dk/faktaomfilm/person/da/7401.aspx?id=7401).

(10). “L’aiguille” at EFG: http://tinyurl.com/l9yp4qg

Original title: “Die Weiche”. Swiss short feature film produced in 1961. It is an unknown film from an unknown director. The EFG portal does not give detailed contextual information about it. Some film scholars think it is an amateur film, which combines different cinematographic techniques in naïve approach too basic for its time (November, 2014, personal communication with Spanish film scholars).

(11). “Metropolis” at EFG: http://tinyurl.com/kmvmylh

Fritz Lang's classic and renowned science fiction film, one of the greatest films of all times. The clip corresponds to the sequence where the robot Maria incites the workers to revolt.

(12). “Vampyr” at EFG: http://tinyurl.com/otunuvv

Also known as “L’etrange aventure de David Gray”, is one of the most known films by Carl Theodor Dreyer and is “one of the founding and defining works of psychological horror cinema” (Rudkin, 2007).

(13). The questionnaire is available at surveys.timelessfuture.com/waisda

(14). The data is made available online in anonymized form at https://github.com/biktorrr/waisda_efg

(15). The “Emolab” project by Frans Hals Museum in Haarlem, The Netherlands. http://www.commit-nl.nl/news/emolab-in-frans-hals-museum

(16). Project “Emotions in Film” at the University of Amsterdam. http://cdh.uva.nl/projects-2012-2013/emotions-in-film/emoties-in-film.html

(17). http://www.fiafnet.org/uk/publications/iifp_subjectHeadings.html

(18). Questions are numbered “q1, q2,…” the full questions are found in the Appendix.

(19). http://www.imdb.com/Sections/Keywords/

References

(All online documents were last accessed on January 26, 2015).

Ådland, M. K., & Lykke, M. (2012). Social Tagging in Support of Cancer Patients' Information Interaction. In *Social Information Research* (Vol. 5, pp. 101–128). UK: Emerald Group Publishing.

Ahn, L. von, & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67. http://doi.org/10.1145/1378704.1378719

Baca, M. (2002). *Introduction to Art Image Access: Issues, Tools, Standards, and Strategies*. Los Angeles, CA: Getty Research Institute.

Bálint, K., & Kovács, A. B. (2012). Focalization and Attachment. Studying the interaction effect of narrative and psychological factors in film viewers' emotional responses. *Pszichológia*, *32*(3), 271–291. http://doi.org/10.1556/Pszicho.32.2012.3.6

Ballan, L., Bertini, M., Del Bimbo, A., Meoni, M., & Serra, G. (2010). Tag Suggestion and Localization in User-generated Videos Based on Social Knowledge. In *Proceedings of Second ACM SIGMM Workshop on Social Media* (pp. 3–8). New York, NY, USA: ACM. http://doi.org/10.1145/1878151.1878155

Ballan, L., Bertini, M., Del Bimbo, A., & Serra, G. (2011). Enriching and Localizing Semantic Tags in Internet Videos. In *Proceedings of the 19th ACM International Conference on Multimedia* (pp. 1541–1544). New York, NY, USA: ACM. http://doi.org/10.1145/2072298.2072060

Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y., & Shachak, A. (2008). Structured versus unstructured tagging: a case study. *Online Information Review*, *32*(5), 635–647. http://doi.org/http://dx.doi.org/10.1108/14684520810914016

Bertini, M., Del Bimbo, A., Ferracani, A., Gelli, F., Maddaluno, D., & Pezzatini, D. (2013a). A Novel Framework for Collaborative Video Recommendation, Interest Discovery and Friendship Suggestion Based on Semantic Profiling. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 451–452). New York, NY, USA: ACM. http://doi.org/10.1145/2502081.2502264

Bertini, M., Del Bimbo, A., Ferracani, A., Gelli, F., Maddaluno, D., & Pezzatini, D. (2013b). Socially-aware Video Recommendation Using Users' Profiles and Crowdsourced Annotations. In *Proceedings of the 2Nd International Workshop on Socially-aware Multimedia* (pp. 13–18). New York, NY, USA: ACM. http://doi.org/10.1145/2509916.2509924

Boer, V. de, Hildebrand, M., Aroyo, L., Leenheer, P. D., Dijkshoorn, C., Tesfa, B., & Schreiber, G. (2012). Nichesourcing: harnessing the power of crowds of experts. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, … N. Hernandez (Eds.), *Ekaw'12: proceedings of the 18th international conference on knowledge engineering and knowledge management* (pp. 16–20). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://doi.org/10.1007/978-3-642-33876-2_3

Bordwell, D., & Thompson, K. (2003). *Film Art: An Introduction* (7th ed.). Mcgraw-Hill College.

Burford, B., Briggs, P., & Eakins, J. P. (2003). A Taxonomy of the Image: On the Classification of Content for Image Retrieval. *Visual Communication*, *2*(2), 123–161. http://doi.org/10.1177/1470357203002002001

Darvish, S., & Chin, A. (2010). Dealing with the Video Tidal Wave: The Relevance of Expertise for Video Tagging. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (pp. 289–290). New York, NY, USA: ACM. http://doi.org/10.1145/1810617.1810679

Eakins, J. P., Briggs, P., & Burford, B. (2004). Image Retrieval Interfaces: A User Perspective. In P. G. B. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, & A. W. M. Smeulders (Eds.), *Image and Video Retrieval* (pp. 628–637). Springer Berlin Heidelberg.

Enser, P. G. B. (2000). Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science*, *26*(4), 199–210. http://doi.org/10.1177/016555150002600401

Fossati, G. (2009). *From Grain to Pixel: The Archival Life of Film in Transition*. Amsterdam University Press.

Freiburg, B., Kamps, J., & Snoek, C. G. M. (2011). Crowdsourcing visual detectors for video search (pp. 913–916). New York, NY, USA: ACM. http://doi.org/10.1145/2072298.2071901

Fu, W.-T., Kannampallil, T., Kang, R., & He, J. (2010). Semantic imitation in social tagging. *ACM Trans. Comput.-Hum. Interact.*, *17*(3), 12:1–12:37. http://doi.org/10.1145/1806923.1806926

Gedikli, F., & Jannach, D. (2013). Improving Recommendation Accuracy Based on Item-specific Tag Preferences. *ACM Trans. Intell. Syst. Technol.*, *4*(1), 11:1–11:19. http://doi.org/10.1145/2414425.2414436

Geisler, G., Willard, G., & Whitworth, E. (2010). Crowdsourcing the indexing of film and television media (pp. 82:1–82:10). Silver Springs, MD, USA: American Society for Information Science. Retrieved from http://dl.acm.org/citation.cfm?id=1920331.1920448

Gibbon, D. C., Liu, Z., Basso, A., & Shahraray, B. (2013). Automated content metadata extraction services based on MPEG standards. *Computer Journal*, *56*(5), 628–645. http://doi.org/10.1093/comjnl/bxs146

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Aroyo, L., & Schreiber, G. (2013). An Evaluation of Labelling-Game Data for Video Retrieval. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, … E. Yilmaz (Eds.), *Advances in Information Retrieval* (pp. 50–61). Springer Berlin Heidelberg.

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Schreiber, G., & Aroyo, L. (2011). On the role of user-generated metadata in audio visual collections (pp. 145–152). New York, NY, USA: ACM. http://doi.org/10.1145/1999676.1999702

Goh, D. H.-L., Ang, R. P., Lee, C. S., & Chua, A. Y. K. (2011). Fight or unite: Investigating game genres for image tagging. *Journal of the American Society for Information Science and Technology*, *62*(7), 1311–1324. http://doi.org/10.1002/asi.21478

Goh, D. H.-L., & Lee, C. S. (2011). Perceptions, quality and motivational needs in image tagging human computation games. *Journal of Information Science*, *37*(5), 515–531. http://doi.org/10.1177/0165551511417786

Golbeck, J., Koepfler, J., & Emmerling, B. (2011). An experimental study of social tagging behavior and image content. *Journal of the American Society for Information Science and Technology*, *62*(9), 1750–1760. http://doi.org/10.1002/asi.21522

Good, B. M., Tennis, J. T., & Wilkinson, M. D. (2009). Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics. *BMC Bioinformatics*, *10*, 313–313. http://doi.org/10.1186/1471-2105-10-313

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up Tags? *D-Lib Magazine*, *12*(1). Retrieved from http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january06/guy/01guy.html

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web* (pp. 211–220). New York, NY, USA: ACM. http://doi.org/10.1145/1242572.1242602

Hildebrand, M., Brinkerink, M., Gligorov, R., van Steenbergen, M., Huijkman, J., & Oomen, J. (2013). Waisda? Video Labeling Game. Presented at the ACM Multimedia, Barcelona.

Hollink, L. (2006). *Semantic annotation for retrieval of visual resources* (Doctoral Dissertation). Vrije Universiteit, Amsterdam. Retrieved from http://hdl.handle.net/1871/10846

Hollink, L., Schreiber, A. T., Wielinga, B. J., & Worring, M. (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies*, *61*(5), 601–626. http://doi.org/10.1016/j.ijhcs.2004.03.002

Huang, C., Fu, T., & Chen, H. (2010). Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, *61*(5), 891–906. http://doi.org/10.1002/asi.21291

Images for the Future. (2009). Waisda? Video Labeling Game: Evaluation Report. Retrieved from http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

Inskip, C., MacFarlane, A., & Rafferty, P. (2008). Content or Context?: Searching for Musical Meaning in Task-based Interactive Information Retrieval. In *Proceedings of the Second International Symposium on Information Interaction in Context* (pp. 72–74). New York, NY, USA: ACM. http://doi.org/10.1145/1414694.1414711

Kang, R., & Fu, W.-T. (2010). Exploratory information search by domain experts and novices. In *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 329–332). New York, NY, USA: ACM. http://doi.org/10.1145/1719970.1720023

Kim, H. H., & Kim, Y. H. (2010). Toward a conceptual framework of key-frame extraction and storyboard display for video summarization. *Journal of the American Society for Information Science and Technology*, *61*(5), 927–939. http://doi.org/10.1002/asi.21317

Klavans, J. L., LaPlante, R., & Golbeck, J. (2013). Subject matter categorization of tags applied to digital images from art museums. *Journal of the American Society for Information Science and Technology*, n/a–n/a. http://doi.org/10.1002/asi.22950

Layne, S. S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging & Classification Quarterly*, *6*(3), 39–62. http://doi.org/10.1300/J104v06n03_04

Lee, C. S., Goh, D. H.-L., Razikin, K., & Chua, A. Y. K. (2009). Tagging, Sharing and the Influence of Personal Experience. *Journal of Digital Information*, *10*(1). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/275

Li, G., Wang, M., Zheng, Y.-T., Li, H., Zha, Z.-J., & Chua, T.-S. (2011). ShotTagger: Tag Location for Internet Videos. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (pp. 37:1–37:8). New York, NY, USA: ACM. http://doi.org/10.1145/1991996.1992033

Lu, C., Park, J., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, *36*(6), 763–779. http://doi.org/10.1177/0165551510386173

Madrigal, A. C. (2014, January). How Netflix Reverse Engineered Hollywood. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/3/

Matusiak, K. K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services*, *22*(4), 283–298. http://doi.org/10.1108/10650750610706998

Melenhorst, M., Grootveld, M., van Setten, M., & Veenstra, M. (2008). Tag-based information retrieval of video content (pp. 31–40). New York, NY, USA: ACM. http://doi.org/10.1145/1453805.1453813

Oomen, J., & Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges (pp. 138–149). New York, NY, USA: ACM. http://doi.org/10.1145/2103354.2103373

Panofsky, E. (1939). *Studies in iconology : Humanistic themes in the art of the Renaissance* (1st Icon ed., 4th print). New York, NY: Harper & Row.

Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0* (1st ed.). De Gruyter.

Rudkin, D. (2007). *Vampyr*. London: University of California Press.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., … Riedl, J. (2006). Tagging, Communities, Vocabulary, Evolution. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (pp. 181–190). New York, NY, USA: ACM. http://doi.org/10.1145/1180875.1180904

Skov, M., & Lykke, M. (2012). Unlocking radio broadcasts: user needs in sound retrieval (pp. 298–301). New York, NY, USA: ACM. http://doi.org/10.1145/2362724.2362779

Smith, G. (2007). *Tagging: People-powered Metadata for the Social Web*. New Riders Press.

Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D., & Zinkham, H. (2008). *For the* *common good: the Library of Congress Flickr pilot project* (Report). D.C.: Government of the United States; Library of Congress. Retrieved from http://www.egov.vic.gov.au/focus-on-countries/north-and-south-america-and-the-caribbean/united-states/government-initiatives-united-states/culture-sport-and-recreation-united-states/libraries-united-states/for-the-common-good-the-library-of-congress-flickr-pilot-project-in-pdf-format-1333kb-.html

Thøgersen, R. (2013). Data Quality in an Output-Agreement Game: A Comparison between Game-Generated Tags and Professional Descriptors. In P. Antunes, M. A. Gerosa, A. Sylvester, J. Vassileva, & G.-J. de Vreede (Eds.), *Collaboration and Technology* (pp. 126–142). Springer Berlin Heidelberg.

Thom-Santelli, J., Cosley, D., & Gay, G. (2010). What do you know?: experts, novices and territoriality in collaborative systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1685–1694). New York, NY, USA: ACM. http://doi.org/10.1145/1753326.1753578

Tirilly, P., Mu, X., Huang, C., Xie, I., Jeong, W., & Zhang, J. (2012). On the consistency and features of image similarity (pp. 164–173). New York, NY, USA: ACM. http://doi.org/10.1145/2362724.2362754

Trant, J. (2009a). Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing Research. *Journal of Digital Information*, *10*(1). Retrieved from http://journals.tdl.org/jodi/article/viewArticle/270

Trant, J. (2009b). *Tagging, Folksonomy and Art Museums: Results of steve.museum's research*. Retrieved from http://www.museumsandtheweb.com/blog/jtrant/stevemuseum_research_report_available_tagging_fo.html

Troncy, R., Huet, B., & Schenk, S. (2011). *Multimedia Semantics: Metadata, Analysis and Interaction*. John Wiley & Sons.

Tsai, L.-C., Hwang, S.-L., & Tang, K.-H. (2011). Analysis of keyword-based tagging behaviors of experts and novices. *Online Information Review*, *35*(2), 272–290. http://doi.org/http://dx.doi.org/10.1108/14684521111128041

Turner, J. M. (2009). Moving image indexing. In *Encyclopedia of Library and Information Sciences* (3rd ed., pp. 3671–3681). New York: Taylor & Francis.

Turner, J. M. (2010). From ABC to http: The Effervescent Evolution of Indexing for Audiovisual Materials. *Cataloging & Classification Quarterly*, *48*(1), 83–93. http://doi.org/10.1080/01639370903341919

Wang, M., Ni, B., Hua, X.-S., & Chua, T.-S. (2012). Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*, *44*(4), 25:1–25:24. http://doi.org/10.1145/2333112.2333120

Westman, S. (2009). Image Users' Needs and Searching Behaviour. In A. Göker & J. Davies (Eds.), *Information Retrieval: Searching in the 21st Century; Human Information Retrieval* (pp. 63–83). John Wiley & Sons, Ltd. Retrieved from http://doi/10.1002/9780470033647.ch4/summary

Wilkie, C. (1999). *Managing Film and Video Collections*. Aslib and Information Management International.

Winget, M. (2009). Describing art: an alternative approach to subject access and interpretation. *Journal of Documentation*, *65*(6), 958–976.

Yeh, M.-C., & Wu, W.-P. (2014). Clustering Faces in Movies Using an Automatically Constructed Social Network. *IEEE MultiMedia*, *21*(2), 22–31. http://doi.org/10.1109/MMUL.2014.24

Zollers, A. (2007). Emerging Motivations for Tagging: Expression, Performance, and Activism. In *Tagging and Metadata for Social Information Organization Workshop, WWW07*.

Group	N	Total tags	Mean	Median	Min	Max	Standard deviation
A. Experts/ No instructions	9	641	71.2	66.0	27	140	40.9
B. Experts/ instructions	9	773	85.89	77.0	48	140	28.17
C. Novices/ No instructions	9	738	82.0	61.0	23	193	58.1
D. Novices / instructions	9	791	87.9	88.0	55	150	31.0

Table 1. Descriptive statistics of number of tags per group (5 film clips, total duration: 700 sec.).

Categories	**A** (Experts/No instructions)	**B** (Experts/Instructions)	**C** (Novices/No instructions)	**D** (Novices/Instructions)
Cinematography	silent film; black and white; fiction	silent film; black and white; close-up	black and white; silent film; drama	black and white; silent film; close-up
Emotions	mystery; danger; fear	danger; help; angry	old; pain; scary	fear; relief; anger
Explanations	rebellion; expressionism; dystopia	expressionism; death; poverty	death; impressionism; luck	lucky; death; menacing music
Facts	door; train; smoking	shadow; smoking; monkey	shadow; workers; train	shadow; monkey; bell
Other	film; dreyer; german	german; vampyr; early cinema	german; vampyr; italy	german; metropolis; french

Table 2. Three most frequent tags in each category of Classification No.1 per group.

	All groups (A, B, C, D)	Experts (No instructions/ Instructions) (A and B)	Novices (No instructions/ Instructions) (C and D)	Experts and Novices (No Instructions) (A and C)	Experts and Novices (Instructions) (B and D)
CINEMATOGRAPHY	0.102	0.340	0.161	0.387	0.024
EMOTIONS	0.001	0.024	0.031	0.003	0.113
EXPLANATIONS	0.338	0.931	0.050	0.136	0.666
FACTS	0.498	1.000	0.190	0.605	0.666
OTHER	0.383	0.222	0.387	0.436	0.546

Table 3. p values from Kruskal-Wallis and Mann–Whitney U test considering the five film clips. Cells in grey scale indicate a statistically significant difference at the p<0.05 level.

Cinematographic tags (sub-type)	Expert tags’ frequencies (Groups A+B)	Novice tags’ frequencies (Groups C+D)
Acting	extras (1); silent film actress (1)	
Copy	restoration (1);	poor picture quality (1)
Editing	rapid cutting (1); parallel cutting (1); reverse (1); editing (1); continuity editing (1);	continuous (1) ; fadeout (1)
Genre	silent film *(mute cinema, mute pictures, silent, silent cinema, silent movie, silent movies)* (25); fiction (4); thriller (3); sound film (2); trailer (2); horror (2); drama (2); documentary feel (1); science fiction (1); melodrama (1)	silent film *(mute cinema, mute pictures, silent, silent cinema, silent movie, silent movies)* (25); fiction (1); thriller (1); horror (1); drama (3)
Mise-en-scene	exterior shots (3); interior shot *(interior scene)* (3); interior (2); decor (1); set design (1); setting (1)	
Narrative	intertitle (7); titles (4); credits (4); intro (2); climax (2); German intertitles (1); end title (1); title card (1); epilogue (1); narrative (1); end (1)	titles (1); end (2); start (1); subtitles (1); sequence (1)
Shot type-framing	close-up (6); long shot (4); high angle (3); camera pan (2); subjective shot (2); shot on location (1); pan shot (1); fear in close-up shot (1); deep focus (1); detail (1); diagonal (1); panning (1); point-of-view (1); crane shot (1); close up interior shots (1); offscreen (1); extreme long shot (1); topshot (1); low angle (1); aerial shot (1)	close-up (5)
Shot-photographic aspects	black-and-white film *(black and white, black & white, black white)* (10); superimposition (3); shadow theatre *(chinese shadows, javanese shadows, shadowplay)* (3); chiaroscuro (1); double exposure (1); vignetting on film (1); tableau (1); trick photography (1); silhuoettes (1); masking (1)	black-and-white film *(black and white, black & white, black white)* (22); shadow theatre *(chinese shadows, javanese shadows, shadowplay)* (1)
Technique-sound	offscreen sound (2) ; scored music (1) ; accompaniment (1) ; musical accompaniment (1)	

Table 4. Cinematographic tags for the five film clips used by experts and novices groups combined (respectively A+B; C+D), including tags in the long-tail portion of the total tags’ distribution.

Category / Group	A Experts /no instructions	B Experts /instructions	C Novices /no instructions	D Novices /instructions	Total
General/Who (e.g., man, bell, dog, animals)	48.16%	40.27%	35.64%	32.59%	38.54%
General/What (e.g., bell ringing, children playing, hug, kissing goodbye)	23.21%	23.03%	21.19%	31.07%	24.88%
Abstract/What (e.g., abandoned, bored, calamity, danger)	15.09%	23.33%	26.37%	27.60%	23.63%
Abstract/Who (e.g., thief, proletarian, friend)	4.84%	7.73%	8.95%	4.99%	6.67%

Table 5. Proportional distribution of Conceptual tags across different categories (Classifications No.3 and 4: the Panofsky/ Shatford matrix) per group. Percentage in relation to the total conceptual tags per group.

Category/Group	A Experts /no instructions	B Experts /instructions	C Novices /no instructions	D Novices /instructions	Total
General	74.76%	66.72%	59.18%	64.63%	65.88%
Abstract	21.19%	31.51%	36.58%	34.40%	31.49%
Specific	4.05%	1.78%	4.24%	0.97%	2.62%

Table 6. Proportional distribution of Conceptual tags across different categories (Classification No.3) per group. Percentage in relation to the total conceptual tags per group.

Groups (n=9)	q18.Perceived usefulness of instructions (categories) (1=not at all useful; 5=extremely useful)			
	Mode	Median	Min	Max
Group A (Experts/no instructions)	2 (n=4)	2	1	5
Group B (Experts/instructions)	3 (n=3) 5 (n=3)	4	1	5
Group C (Novices/no instructions)	3 (n=6)	3	1	5
Group D (Novices/instructions)	3 (n=4) 4 (n=4)	4	3	5

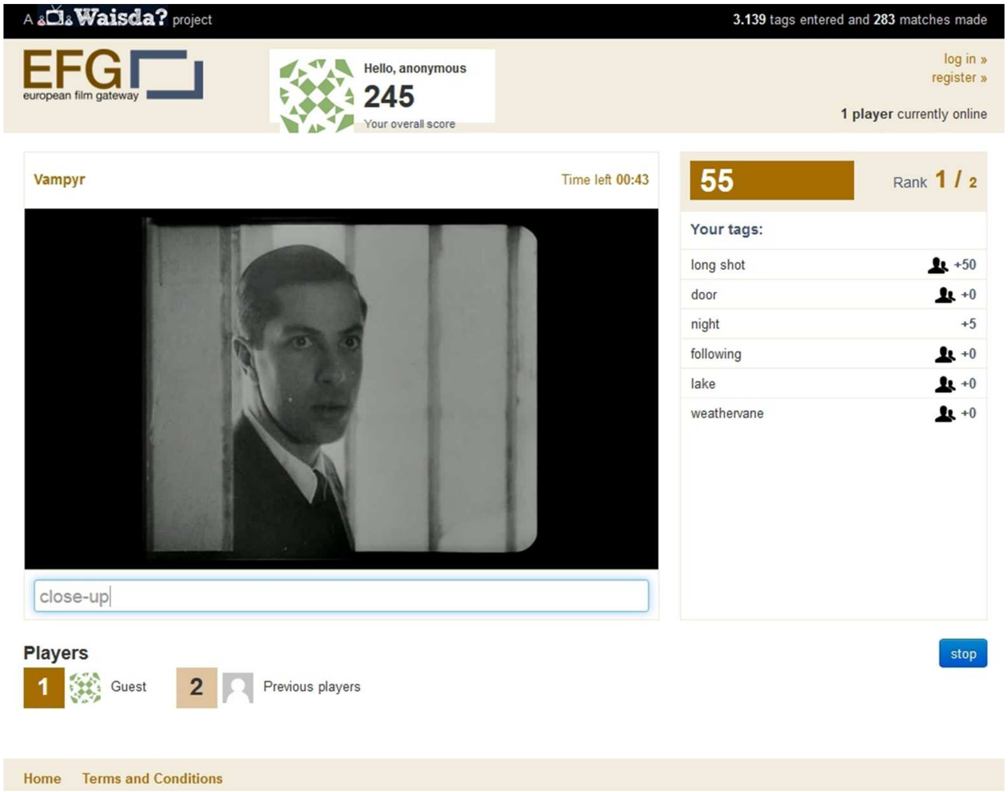
Table 7. Frequencies of ranking on a 5 point Likert scale the usefulness of instructions during tagging (1=not at all; 5=extremely).

Groups (n=9)	q12.Difficulty in coming up with tags				q13.Possibility of entering all tags				q15.Influence of scoring in game motivation			
	Mode	Median	Min	Max	Mode	Median	Min	Max	Mode	Median	Min	Max
Group A (Experts/no instructions)	2 (n=3); 4 (n=3)	4	2	5	4 (n=5)	4	1	5	1 (n=3) 2 (n=3)	2	1	5
Group B (Experts/instructions)	3 (n=3) 4 (n=3)	3	2	5	2 (n=3) 3 (n=3)	3	2	5	4 (n=3)	3	1	5
Group C (Novices/no instructions)	2 (n=4)	3	2	5	4 (n=4)	4	3	5	1; 2; 4; 5 (n=2)	3	1	5
Group D (Novices/instructions)	2 (n=3); 3 (n=3)	3	2	5	4 (n=4)	4	1	5	4 (n=3)	4	1	5

Table 8. Frequencies of ranking on a 5 point Likert scale different aspects of tagging behavior: (q12: 1=very difficult; 5=very easy); (q13: 1=not possible; 5= possible); (q15: 1=not at all influential; 5=extremely influential).

Groups (n=9)	q20.Perceived usefulness of entered tags (No=0 / Uncertain=1 /Yes=2)	
	Mode	Median
Group A (Experts/no instructions)	1 (n=4)	1
Group B (Experts/instructions)	1 (n=6)	1
Group C (Novices/no instructions)	2 (n=8)	2
Group D (Novices/instructions)	2 (n=6)	2

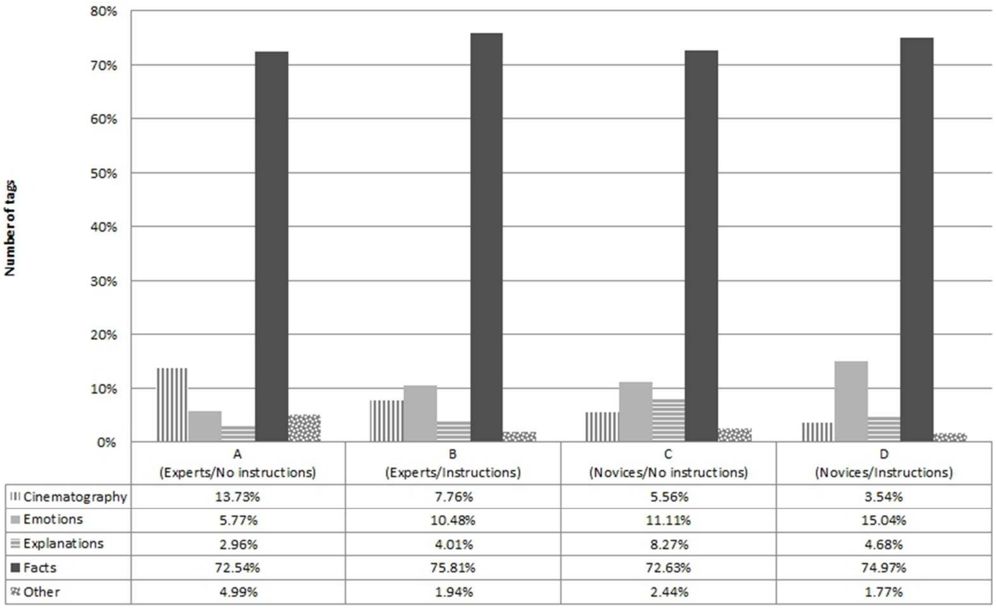
Table 9. Frequencies of ranking previous knowledge of the experiment films, on a 3 point Likert scale: 0=no previously seen and no knowledge; 1=either seen or some knowledge; 2=previously seen and had knowledge.



close-up

long shot	+50
door	+0
night	+5
following	+0
lake	+0
weathervane	+0

344x270mm (72 x 72 DPI)



	A (Experts/No instructions)	B (Experts/Instructions)	C (Novices/No instructions)	D (Novices/Instructions)
▥ Cinematography	13.73%	7.76%	5.56%	3.54%
■ Emotions	5.77%	10.48%	11.11%	15.04%
▤ Explanations	2.96%	4.01%	8.27%	4.68%
■ Facts	72.54%	75.81%	72.63%	74.97%
▨ Other	4.99%	1.94%	2.44%	1.77%

306x186mm (72 x 72 DPI)