Data Quality in an Output-Agreement Game: A Comparison between Game-Generated Tags and Professional Descriptors

Rasmus Thogersen

Victoria University of Wellington, New Zealand
Rasmus.thogersen@vuw.ac.nz

Abstract. A novel way to address the challenge of creating descriptive metadata for visual cultural heritage is to invite users to play Human Computation Games (HCG). This study presents an investigation into tags generated by an HCG launched at The Royal Library of Denmark and compares them to descriptors assigned to the same images by professional indexers from the same institution. The analysis is done by classifying tags and descriptors by term-category and by measuring semantic overlap between the tags and the descriptors. The semantic overlap was established with thesaurus relations between a sample of tags and descriptors.

The analysis shows that more than half of the validated tags had some thesaurus relation to a descriptor added by a professional indexer. Approximately 60% of the thesaurus relations were either 'same/equivalent' and roughly 20% were 'associative' and 20% 'hierarchical'. For the hierarchical thesaurus relations it was found that tags typically describe images at a less specific level than descriptors.

Furthermore game-generated tags tend to describe 'artifacts/objects' and thus typically represent what is in the picture, rather than what it is about. Descriptors also primarily belonged to this term-category but also had a substantial amount of 'Proper nouns', mainly named locations. Tags generated by the game, not validated by player-agreement, had a much higher frequency of 'subjective/narrative' tags, but also more errors and a few cases of vandalism. The overall findings suggest that game-generated tags could complement existing metadata and be integrated into existing workflows.

Keywords: Games with a purpose, crowdsourcing, image indexing, cultural heritage institutions, participatory cultural heritage, Output-agreement games.

1 Introduction

This paper aims to evaluate the outcome of the crowdsourcing tool Games with a Purpose (GWAP) or Human Computation Games (HCG) against professionally created metadata. It describes the Royal Library of Denmark's use of an Output-agreement game to index 2079 photographs in 2010 and how the metadata output created via the game compares to the metadata already created in-house at the Royal

P. Antunes et al. (Eds.): CRIWG 2013, LNCS 8224, pp. 126–142, 2013.

Library. As crowdsourcing is becoming a part of the common toolkit in the cultural heritage sector, an understanding of how the tags can complement traditional knowledge organization systems is needed. While numerous studies have investigated the relation between tags, to the best of my knowledge no previous studies have investigated the relationship between the output of a game and professional index terms.

2 Background

As shown in Figure 1, making cultural heritage digital can be viewed as a 7-step process [1]. In 2010, the total cost of digitizing the content of Europe's cultural heritage institutions (Libraries, Archives and Museums) was estimated to be approximately 100 Billion Euro, which only covers the cost of Selecting, Creating, Describing, Managing and Preserving [2]. This paper covers the task of Describing.

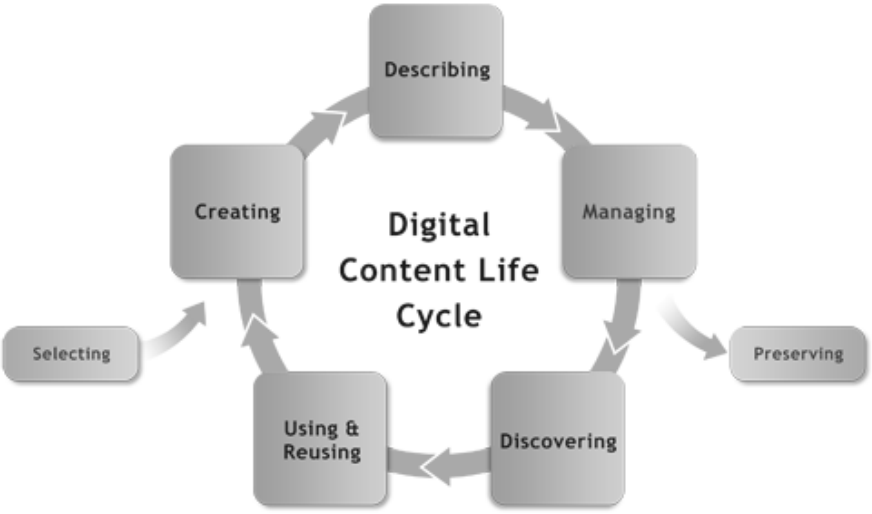

Fig. 1. Digital Content Life Cycle (Source: DigitalNZ)

Describing is mainly a matter of surrogacy i.e. creating data about the content, also known as metadata. The report estimates the cost of metadata-creation to range between 3.5-15 Euro for each object, depending on the state of the object, the type and the organizational context. This covers technical metadata (e.g. file-type, checksum), administrative metadata (e.g. copyright, provenance) and descriptive metadata (e.g. author, title and subject) the latter of which cannot always easily be ascertain via automatic means and often requires human interpretation to assess. The presence of subject metadata (keywords) is essential for content discovery via searching or browsing and represents of one of the challenges facing cultural heritage

institutions when migrating into a digital environment: the creation of subject metadata for the rapidly increasing amount of content.

An increasingly popular approach is to rely on user-created index terms, typically by allowing/inviting the users to tag directly in the online catalogs or by publishing the content on external content aggregators with a preexisting social infrastructure already in place (e.g. Flickr or LibraryThing). Both approaches are variations of crowdsourcing tools and make particular sense in the realm of digital image collections in the cultural heritage sector for two reasons:

- Cultural heritage institutions i.e. galleries, archives, museums and libraries have historically been relying on volunteerism [3] and crowdsourcing is a natural extension of this notion.
- Image materials are notoriously hard to index, which is reflected in the literature, to the extent that a more user-driven and 'democratic approach' to image indexing was proposed in 1996 [4] – a decade before the term crowdsourcing was coined [5].

Crowdsourcing in a cultural heritage context can serve multiple purposes. Aside from the rationalization/cost and how the crowd can accomplish things single indexers/institutions cannot - there is another benefit in engaging patrons in some sort of activity, be it describing, digitizing or even co-creating the collection; it can be seen as marketing/dissemination of the library resources. The activities can stimulate interest and lead to discovery and the very notion of inviting the wider public to collaborate is a way for the institution to signal openness and approachability.

One concern, however, when engaging in any kind of crowdsourcing project is the behavior of the eponymous 'crowd'. Cultural heritage institutions have relied on volunteering, but another value embedded in the profession is the notion of authoritative delivery of high quality and un-biased information [6] - an ideal that can be hard to uphold if the institution itself isn't in control of the content it provides. Lascarides states that digital vandalism in crowdsourcing is far rarer than most people expect, but does also note that given the novelty of the field, precious little is actually known about the quality of the output of crowdsourcing projects [7]. An alternative method to tagging only recently applied to image collections in the cultural heritage sector, is HCG, a crowdsourcing tool that uses gamification in the indexing process and relies on user-agreement to create validated tags.

This work aims to investigate the output of an HCG by comparing the user-generated keywords (Tags) to professionally assigned keywords (Descriptors) to deepen our understanding of its feasibility in the cultural heritage sector and is carried out using data from an HCG called 'Make a Difference'[1] developed at The Royal Library in Copenhagen, Denmark and aims to answer the following questions:

RQ1: How similar are the tags of an output-agreement game to the descriptors provided by professional indexers? *Similarity is defined as the thesaurus-derived relations (and strength of those relationships), as the semantic overlap between the*

[1] Translated from Danish by the author.

two kinds of keywords should provide an estimation of the quality of the tags by using the descriptors as a gold standard set.

RQ2: What is the difference in the term-type of the labels assigned by gamers and indexers respectively? *To successfully utilize game-generated tags and how they can complement descriptors, a better understanding of their characteristics are needed.*

3 Related Literature

This section presents the context in which the study takes place. First describing the problems of assigning keywords to images and then introducing crowdsourcing in the cultural heritage context, followed by a description of Human Computation Games in general and the specific type of game created by the Royal Library, Output-agreement games.

Image Indexing is divided into two broad concepts: 'Content-Based Image Indexing' and 'Concept-Based Image Indexing'. The former relates to the picture 'as is it' and refers to computational methods in which a software application decodes an image and returns descriptors [8]. This might be easy for colors or simple patterns, but moving beyond pre-iconographic descriptions presents a computer with significant problems, e.g. describing a mood, identifying a location or interpreting a meaning [6], which is why the reliance on 'Concept Based Image Indexing' still is relevant.

'Concept-Based Image Indexing' presents human indexers with its own unique challenges, as they attempt 'to translate visually coded knowledge into a verbal surrogate' [9]. Indexing images with verbal descriptions is likely to be more subjective than it is when indexing texts [10]. This knowledge led researchers to suggest a 'democratic' approach to image indexing in which users, not indexers, provide the keywords [4]. This was a precursor to the now-widespread phenomenon folksonomies, which is the non-controlled, bottom-up vocabulary that emerges when users tag objects via collaborative information services, such as Flickr, delicious or LibraryThing.

Crowdsourcing is a relatively new concept and is a sort of umbrella term for various practices that involve mass-collaboration on online platforms. The term itself was coined by Howe in his seminal 2006 article in Wired Magazine [5], in which he describes how companies can reduce costs dramatically by outsourcing certain processes to the crowd, rather than having highly trained (and thus costly) professionals perform menial tasks. The approach is highly adaptable, which invariably leads to a plethora of use-cases and makes any attempts at a definition and construction of taxonomies more of an ongoing conversation [11].

In the cultural heritage sector, crowdsourcing is used as a way to collaborate with users via social media platforms, typically centered around a certain collection; has been utilized for correction, contextualization, co-curation, complementing,

crowdfunding and classification [12]; and was heralded by Holley [13] as a highly promising approach to problem-solving for libraries in general and image collections in particular. Often, the publication of digitized images is delayed, not because of technical issues such as scanning or publishing, but by the lack of metadata to make the images retrievable via browsing/searching.

Studies of crowdsourcing participants have revealed that they are motivated to work either for extrinsic reasons (monetary rewards, learning new skills or recognition from external parties) or intrinsic reasons (partaking in a community or enjoyment) [14]. When deciding on a platform for a crowdsourcing project, these should be taken into account and since monetary rewards aren't likely to be one of the motivational factors, crowdsourcing projects in the cultural heritage sector should aim to either teach the participants something along the way, build a sense of community or make the experience as fun and enjoyable as possible.

Human Computation Games is a method pioneered by Louis von Ahn and Laura Dabbish in 2004 with the ESP-game, as a way to address the image labeling challenge, i.e. describing the deluge of images available online - both digitized and born digital materials. Rather than relying on content based image retrieval, which doesn't work well enough [15], they came up with the novel idea of getting people do it for free, by taking advantage of their desire to be entertained, and launched it under the moniker Games With A Purpose (GWAP).

The ESP-game is a browser-based game in which two anonymous players are connected and shown the same image. Each of them is then tasked with assigning labels to the image and guessing the labels of the other player. A successful match scores both players points. This part of the game – obtaining multiple labels – has since then become one of the established ways to ensure quality in crowdsourcing projects, i.e. by some sort of defensive design [16]. The validation threshold, i.e. the number of players that have to agree on a term, can be modified according to local preferences.

Figure 2 provides an example from the 'Make a Difference' game where X=3, as three players had to agree on a tag. Unlike the ESP-game, where gameplay is simultaneous, play-sessions can be asynchronous. The three sets of tags from the three players can be added over time, and once three players agree on a given term, it becomes valid. The idea of validation ties into the second innovative gameplay-component from the ESP-game: the idea of Taboo-words. Once a label is validated, it appears in all subsequent games on the screen in grey and the game will no longer accept this exact label, effectively forcing players to provide labels beyond the most obvious ones and makes the indexing process an iterative one as the image runs through multiple play-sessions. In Figure 2, for example, two Taboo words already exist as the three players play the game. Each of the players adds 6 tags to the image, one of which all three agree on ('Statue'). That term then gets transferred to the Taboo-words. Each of the Free Tags and 2Vtags are stored and, should the next player add either 'Summer', 'Sky' or 'Boat', they will also become valid and, therefore, Taboo words.

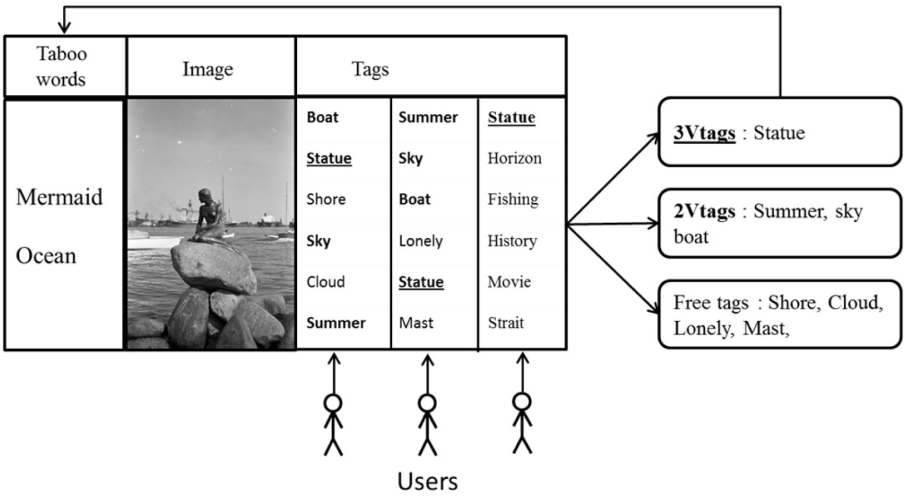

Fig. 2. Example of the Output-agreement gameplay from Make a Difference

Since the inception of the ESP-game, the GWAP platform has developed into a sort of running experiment in HCG, with a wide range of games and media types involved [17]. They are typically very simple, fast and intuitive and can be separated into four fundamental classes [18]:

- **Output-agreement Games**
  All players are given the same input and must produce outputs based on the common input.
- **Input-agreement Games**
  All players are given inputs that are known by the game (but not by the players) to be the same or different. The players are instructed to produce outputs describing their input, so their partners are able to assess whether their inputs are the same or different. Players see only each other's outputs.
- **Inversion-problem Games**
  The first player has access to the whole problem and gives hints to the second player to make a guess. If the second player is able to guess the secret, we assume that the hints given by the first player are correct.
- **Output-optimization Games**
  All players are given the same input and their outputs are the hints of other players' outputs.

The ESP-game as well as the 'Make a Difference'-game investigated in this paper are both Output-agreement games.

Use of HCG as a crowdsourcing tool is not yet a widespread practice in the cultural heritage sector, which can probably be attributed to the fact that developing a game in-house, until very recently, required specialized knowledge. Some examples are the OCR-correction game DigitalKoot from the Finnish National Library as well as the

Dutch 'Waisda?' an Output-agreement game for audiovisual materials. The recent publication of the open source software suite Metadata Games[2], which aims to facilitate local implementation of HCG and is especially targeted towards the humanities, makes evaluation and research into the application of games to create new metadata and complement existing institutional metadata more relevant than ever.

4 Data Collection

In November 2010 The Royal Library of Denmark launched the Output-agreement Game 'Make a Difference' via the social software Facebook, with the stated dual purpose of describing the Danish cultural heritage and collecting money for Save the Children – Denmark. Inspired by the ESP-game, a recently digitized collection of 2079 photographs by the Danish photographer Sven Türck were uploaded, and the crowd was invited to tag the images [19]. For each validated tag (X=3) a donation of 2 DKK was given up to a total of 5000 DKK (provided by external funding). In total, 235 users logged into the game during the ca. two weeks it was open, and they provided a total of 22787 tags, of which 2516 were validated.

The Sven Türck collection had previously been published online by The Royal Library, and the images were already classified by professional indexers to facilitate browsing/searching. As both the professional and the gamer perspective existed, the data generated by the game was suitable for this sort of investigation. The Descriptors were obtained directly from the photo archive via The Royal Library's OAI-server as MODS XML-files, and the game-generated tags were supplied by the developers of the game. The tags were divided into three categories – the non-validated tags (Free Tags), tags validated by two players (2Vtags) and tags validated by three players (3Vtags).

Table 1. Total number of terms for 2079 images

	Free tags	2Vtags	3Vtags	Descriptors
	22787	4743	2516	7306[3]
Average	11	2.3	1.2	3.5

[2] http://metadatagames.com/about/

[3] 1950 of the 2079 images contain the Descriptor 'Denmark'. This descriptor is seemingly a prerequisite for adding any location metadata in the system, more than an actual conscious decision from the indexer and is omitted for the analysis. 'Denmark' is meaningless as a search term; as it will result in almost total recall of the entire collection, it does not have any discriminatory power. In order to normalize the data and prepare it for automated analysis, compound descriptors with two words (omitting proper nouns) were split into separate descriptors and subsequently treated as such.

5 Research Design

5.1 Semantic Overlap

To determine similarity, the simplest approach is to look at syntactic overlap, which relies on character-for-character analysis and determining overlap on a completely binary scale. An extension of this is fuzzy matching, an approach that takes orthographical (e.g. British and American spelling) and morphological (e.g. singular and plural) variations into account and can be automated by a stemming algorithm. To deepen the understanding of the relation between the two types of keywords, the scope can be widened by introducing 'related meaning' using the knowledge structure from a thesaurus.

The method was pioneered by Voorbij [20] and was originally used as a way to determine similarity between title keywords and subject descriptors in the OPAC of the National Library of the Netherlands; titles and keywords from 475 records were scrutinized by subject librarians and assigned a score from 1-7, depending on how similar the keyword was to the title. The method was adapted and modified by Kipp [21] to determine similarity between keywords assigned by authors, indexers and taggers, respectively. Since then, the Voorbij/Kipp approach has been used/adapted by the original authors [22-23] and other researchers [24-27]. While each of these studies represent slightly different approaches, the common idea is to categorize term relations according to the knowledge structure from a thesaurus to determine a semantic overlap. The studies in which term comparisons have been done usually use the formal ontology of the descriptors as a 'reference standard' allowing for a certain amount of automatic analysis, e.g. if a tag has a formal associative relation to a descriptor according to Library of Congress Subject Headings, the relation is established, but a looser interpretation of 'associative' has also been adopted [24-25]. Without a reference standard – as was the case in this study[4] - one can either opt for a more exclusive approach in which the associative relations are ignored altogether or choose some external resource as a standard for comparison. As the analysis would be poorer without connecting obvious semantic dots such as 'fisherman' and 'fishing' an external source for comparison between tags and descriptors was chosen. To ensure rigor in the analysis, the Danish lexical-semantic database DanNet[5] was used in cases of doubt to establish the associative relation.

Standard guides for constructing thesauri define three overarching types of relationships, expressed at various levels of granularity:

- **Equivalence** (Same, Equivalence)
- **Hierarchical** (Narrower, Broader, Part-Whole, Whole-Part, Literal-Descriptor, Tag-Literal)
- **Associative** (Associative)

These relationships can then be ranked according to their strength. The concept of semantic strength was introduced by [25] as a way to do exclusive coding of semantic relations:

[4] The Descriptors are not assigned from a controlled set of subject headings, but chosen ad hoc
[5] In particular the visualisation tool of the dataset published at `andreord.dk`

1. **Same**
   A syntactic match between Tag and Descriptor
2. **Equivalence**
   Tag and Descriptor denote identical concept, i.e. synonyms
3. **Narrower Term**
   Tag is more specific than Descriptor e.g. 'villa – house'
4. **Broader Term**
   Tag is less specific than Descriptor e.g. 'sport – soccer'
5. **Part-Whole**
   Tag describes a more specific part of the Descriptor e.g. 'door – house'
6. **Whole-Part**
   Tag describes a term of which the descriptor is part e.g. 'beach – sand'
7. **Literal-Descriptor**
   Tag is a proper noun for an abstract Descriptor e.g. 'street – Bunny Street'
8. **Tag-literal**
   Tag is an abstract term for a proper noun Descriptor 'lady – Queen Margrethe II of Denmark'
9. **Associative**
   Tag has a direct relation to Descriptor according to DanNet, but not one covered by relation 1-8.

Analysis
Due to time-constraints, a subset of the images (n=320) was chosen randomly for analysis. Each tag was compared to the entire set of descriptors assigned to the same image. Coding was done exclusively, only allowing for one relation to be assigned to each tag and always assigning the strongest semantic relation identified.

Table 2. Number of terms in sample (n=320)

	Free tags	2Vtags	3Vtags	Descriptors
Total	2480	746	380	1112

The total semantic overlap is used to determine the similarity between the set of tags and the set of descriptors and is expressed by the frequency of overlap between the two.

5.2 Term-Categories

In order to code the Descriptors and Tags, the unique values from each dataset were extracted to express the vernacular vocabularies of the different datasets.

Table 3. Number of unique terms on vocabulary level

	Free tags	2Vtags	3Vtags	Descriptors
Total	4121	1040	600	905

Preliminary categories, informed by related literature [28-30] on image indexing, were constructed. The crystallization of the final categories however, was the result of an iterative process i.e. they were continually modified during the immersion in the data. No consensus exists among the creators of these frameworks, although some ideas are ubiquitous: Object, event, location, time and interpretation. These informed the initial term-categories:

- **Artifact/object**
  Static objects in the image, e.g. nouns like *man, table, boat, beach.* These terms refer to general things seen *in* the image or its *ofness.*
- **Action/event**
  Something 'happening', e.g. *dinner, gathering, jumping.*
- **Proper Noun**
  Named places, object or people, e.g. *Copenhagen, The Little Mermaid, Ingrid (1910-2000) droning.*
- **Subjective/Narrative**
  Narrating or interpreting terms, e.g. *idyllic, boring, loving*. These terms attempt to express what the picture is *about.*
- **Time**
  Words describing time, e.g. *winter, evening, October*
- **Errors**
  Spelling mistakes and typos. Not a term category per se, but nonetheless worth measuring considering the uncontrolled nature of tags.

These were later supplemented by three other emerging categories found during the first analysis of the Free tags.

- **Modern**
  Slang or neologisms, often in English e.g. *hot, cool, nice, skyline*
- **From Image**
  In a few cases, seemingly non-sense words are lifted directly from the picture, typically from a sign in the image, such as the name of a shop, e.g. '*NEYE*' or 'K133'. This was the only term-category requiring validation by looking at the image.
- **Obscene**
  Malicious tags or vandalism.

'Make a Difference' technically allowed for multiple-word tagging of the images, so a number of compound tags were observed. As multiple-word tagging is useful for Proper Nouns, e.g. 'Frederiksborg Castle' or 'University of Copenhagen', or qualifying tags, e.g. 'Fast car', this option made sense, but also resulted in different kinds of compound tags, not belonging to either of those categories. These compound tags were initially isolated and then subjected to a refinement; four different subcategories of Compound terms were identified and mapped to the overall categories.

- **Two-Term Concepts**
  e.g. 'Flora_danica' or 'fishing_net'. These are counted as 'Artifacts/objects'.
- **Refining Tags**
  Tags which describe another tag in detail by serving as a qualifier, i.e. adjective-noun pairs like 'old_man' or 'short_hair'. These are counted as 'Subjective/narrative'.
- **Title Tags**
  Narrative string of tags, often explaining the situations depicted. Examples would be either 'reading over the shoulder' or 'dairyman shows the children the butterchurn, it is a jar of butter'. These are counted as 'Subjective/narrative'.
- **Multiple Concept-Tags**
  Strings of unrelated tags, usually comma-separated like 'boys, nature' or 'farm, trees, building, winter'. These are counted as 'Errors'.

Analysis
The term-category analysis was done by listing all tags in a spreadsheet and assigning each tag one of the term-categories described above. In cases of doubt (e.g. the From Image category) the actual images were consulted, but in most cases only the tags were considered.

6 Findings

The Semantic Overlap found between the different categories of tags and the Descriptors is listed in Tables 4-6.

Table 4. Thesaurus relations between Free tags and descriptors

	Free tags (n=2480)		
Relation type	Frequency	% of Total semantic overlap	*M (SD)*
Same (syntactic match)	365	40.24 %	1.12 (1.12)
Equivalence	37	4.08 %	0.11 (0.37)
Narrower	54	5.95 %	0.17 (0.49)
Broader	74	8.16 %	0.23 (0.54)
Part-Whole	9	0.99 %	0.03 (0.16)
Whole-Part	53	5.84 %	0.16 (0.48)
Literal-descriptor	13	1.43 %	0.04 (0.25)
Tag-literal	52	5.73 %	0.16 (0.47)
Associative	250	27.56 %	0.77 (1.34)
Total semantic overlap	907	100 %	0.36 (0.48)

Table 5. Thesaurus relations between 2Vtags and descriptors

	2Vtags (n=746)		
Relation type	Frequency	% of Total semantic overlap	*M (SD)*
Same (syntactic match)	205	54.52%	0.7 (0.78)
Equivalence	12	3.19%	0.04 (0.2)
Narrower	11	2.93%	0.04 (0.19)
Broader	33	8.78%	0.11 (0.39)
Part-Whole	6	1.6%	0 (0.06)
Whole-Part	13	3.46%	0.04 (0.21)
Literal-descriptor	2	0.53%	0.01 (0.08)
Tag-literal	20	5.32%	0.07 (0.28)
Associative	74	19.68%	0.25 (0.61)
Total semantic overlap	376	100%	0.50 (0.50)

Table 6. Thesaurus relations between 3Vtags and descriptors

	3Vtags (n=380)		
Relation type	Frequency	% of Total semantic overlap	*M (SD)*
Same (syntactic match)	132	61.68 %	0.56 (0.65)
Equivalence	5	2.34 %	0.02 (0.14)
Narrower	7	3.27 %	0.03 (0.17)
Broader	17	7.94 %	0.07 (0.28)
Part-Whole	2	0.93 %	0.01 (0.09)
Whole-Part	3	1.40 %	0.01 (0.11)
Literal-descriptor	2	0.93 %	0.01 (0.09)
Tag-literal	8	3.74 %	0.03 (0.18)
Associative	40	18.69 %	0.17 (0.4)
Total semantic overlap	214	100 %	0.56 (0.49)

Overall, the findings suggest that the method of doing semantic comparison yields richer results than merely doing syntactic analysis when comparing metadata for images, as the overlap increased significantly with the inclusion of the thesaurus relations. Even though the players of the game might not use the exact same terms as the professional indexers, there is still a significant overlap in what they see in the picture.

The tags with hierarchical relations were overall on a higher level of abstraction (Broader, Whole-Part and Tag-literal) in the sample. The Free tags had the largest proportion of associative relations and fewer syntactic matches than the validated tags.

Table 7. Percentage of tags with thesaurus relations with descriptors

	Free tags (n=2480)	2Vtags (n=746)	3Vtags (n=380)
Frequency of semantic overlap (%)	907 (36.57%)	376 (50.40%)	214 (56.31%)

As seen in Table 7, more than half of all validated tags and more than a third of the Free tags had some sort of semantic relation to the Descriptors, predominantly the 'Same'-relation. Image indexing being complicated [9], the total semantic overlap must be considered substantial.

The Term-category analysis was initially done on vocabulary level, i.e. the unique terms (Table 8), and the total distribution for all tags and Descriptors was then extrapolated (Table 9).

Table 8. Term-category distribution among unique terms

Category	Free Tags	2Vtags	3Vtags	Descriptors
Artifacts/objects	2345 (56.9%)	829 (79.7%)	505 (84.2%)	469 (51.8%)
Actions/events	392 (9.5%)	82 (7.9%)	45 (7.5%)	31 (3.4%)
Proper noun	316 (7.7%)	91 (8.8%)	41 (6.8%)	382 (42.2%)
Subjective/narrative	380 (9.2%)	21 (2%)	3 (0.5%)	6 (0.7%)
Modern	50 (1.2%)	3 (0.3%)	1 (0.2%)	0 (0%)
From image	11 (0.3%)	1 (0.1%)	0 (0%)	0 (0%)
Time	34 (0.8%)	4 (0.4%)	2 (0.3%)	5 (0.6%)
Error	575 (14%)	9 (0.9%)	3 (0.5%)	12 (1.3%)
Obscene	18 (0.4%)	0 (0%)	0 (0%)	0 (0%)
Total	4121	1040	600	905

Looking at the distribution among non-unique terms, almost 80% of the Free Tags and almost 90% of the 2Vtags and 3Vtags were found to be 'Artifact/objects' - by far the most frequent type of term category observed. The game is set up reward players that guess other players' guesses, so it is not surprising that most players tag what is *in* the picture, rather than what it is *about*, since this is a logical game-play strategy to maximize your score.

The frequency of 'Proper nouns' is stable across all three levels of validation for the Tags. These are typically very recognizable Danish landmarks, e.g. Copenhagen City hall or the Statue of the Little Mermaid. There is a substantially higher ratio of 'Proper Nouns' in the Descriptors. This information can take time and research (beyond looking the photograph) to determine and is therefore less suitable for a fast-paced tagging game. The validation process works, as the error-rate drops from 4.9% to 0.63% for the 2Vtags and further down to 0.5% for the 3Vtags, which interestingly is very close to the 0.3% for the Descriptors and clearly demonstrates the immediate advantage of HCG.

Table 9. Term-category distribution among non-unique terms

	Free tags (*n*=2079)[6]		2Vtags (*n*=1881)	
Category	Frequency (%)	*M(SD)*	Frequency (%)	*M(SD)*
Artifacts/objects	12271 (79%)	5.9 (2.93)	4185 (88.24%)	2.22 (1.3)
Actions/events	831 (5.4%)	0.4 (0.86)	185 (3.9%)	0.1 (0.35)
Proper nouns	909 (5.9%)	0.44 (0.82)	288 (6.07%)	0.15 (0.39)
Subjective/narrative	583 (3.8%)	0.28 (0.61)	39 (0.82%)	0.02 (0.15)
Modern	56 (0.4%)	0.03 (0.17)	3 (0.06%)	0 (0.04)
From image	13 (0.1%)	0 (0.06)	1 (0.02%)	0 (0.02)
Time	71 (0.5%)	0.03 (0.19)	12 (0.25%)	0.01 (0.08)
Errors	762 (4.9%)	0.37 (0.64)	30 (0.63%)	0.02 (0.13)
Obscene	30 (0.2%)	0.01 (0.12)	0 (0%)	0 (0)
Total	15525 (100%)	*7.46 (6.4)*	4743 (100%)	*2.52 (1.33)*

	3Vtags (*n*=1517)		Descriptors (*n*=2062)	
Category	Frequency (%)	*M(SD)*	Frequency (%)	*M(SD)*
Artifacts/objects	2245 (89.2%)	1.48 (0.86)	4479 (61.3%)	2.17 (1.77)
Actions/events	97 (3.9%)	0.06 (0.28)	590 (8.1%)	0.29 (0.59)
Proper nouns	149 (5.9%)	0.1 (0.3)	2062 (28.2%)	1 (1)
Subjective/narrative	8 (0.3%)	0 (0.06)	117 (1.6%)	0.06 (0.23)
Modern	1 (0%)	0 (0.03)	0 (0%)	0 (0)
From image	0 (0%)	0 (0)	0 (0%)	0 (0)
Time	3 (0.1%)	0 (0.04)	37 (0.5%)	0.02 (0.13)
Errors	13 (0.5%)	0.01 (0.09)	21 (0.3%)	0.01 (0.1)
Obscene	0 (0%)	0 (0)	0 (0%)	0 (0)
Total	2516 (100%)	*1.66 (0.87)*	7306 (100%)	*3.54 (1.96)*

A total of 30 Obscene Free tags were found, which shows that vandalism does happen. Most of these were profanity, but a very few cases were racial and sexual slur, which could offend and hurt the users of the collections. These were naturally weeded out by the validation process, but the presence of obscene words in such a short-lived and altruistic project, does demonstrate that vandalism will occur eventually and that we cannot blindly trust the crowd to always have the best intentions.

[6] *n* denotes the number of images in which the tags/descriptors occurred.

Aside from cleaning the metadata, the validation process also cuts off 'the long tail' of the dataset, i.e. the marginal expressions and subjective observations not likely to be echoed by another player. The Subjective/narrative, Modern, From image and Time term-category are hardly represented in the 2Vtags or 3Vtags. One of the strengths of the folksonomy is that it can express a multitude of interpretations and viewpoints, an Output-agreement Game with a validation threshold is clearly not be the best way to accumulate these types of tags.

7 Discussion and Outlook

Using thesaurus relations, it was shown than more than half of the validated tags (both 2Vtags and 3Vtags) had some sort of semantic relation to the Descriptors. Considering the complicated nature of assigning keywords to images, this overlap lends credibility to the overall quality of the tags to warrant implementation into the catalog to some extent.

In this case, the validation process prevented errors and the few cases of vandalism. As the errors in the 2Vtags are only slightly more frequent than the errors in the Descriptors, one recommendation would be to set the validations threshold to 2 rather than 3 as it was in Make a Difference, providing almost twice as many tags as access points. An even more radical approach would be to simply use all Free Tags generated in true 'democratic' [1] fashion. Circumventing the validation process entirely will result in a much higher number of tags, but also introduce flaws in the catalog, the most prevalent of these being simple typing mistakes or common spelling errors, but also possible obscene tags. While extremely rare, they are in themselves enough to argue against a completely open policy in which every contribution by the crowd should be considered equal. There are two ways to deal with this problem:

Pre-tag screening would entail a mechanism of auto-correction, based on either a dictionary or some existing taxonomy that only allows certain terms to be entered, which might rob the final outcome of some of the more creative tags.

Post-tag screening would happen on vocabulary level rather than object level and would take place at regular intervals before allowing the tags to be introduced as proper metadata in the catalog. Catalogers wouldn't need to verify images, but simply scan word-lists for errors and obscenity.

Almost 90% of the 2Vtags and 3Vtags belong to the 'Artifacts/objects' term category. This is hardly surprising considering the nature of Output-agreement games; as the gameplay rewards users for guessing what other people see in the image, the most efficient and obvious strategy is to describe what is *in* the picture. The term-category 'Proper nouns' wasn't very prevalent in the tags, but it features much more prominently in the descriptors. One possible combination of the two types of keywords would be to let the indexers add 'Proper nouns' (mainly locations and personal names) and let the players add information about 'Artifacts/objects', as the game lends itself well to those sorts of descriptions.

Make a Difference was only open to the public for a short time, as the goal was to reach approx. 2500 3Vtags. Having an average of just 1.2 validated tags for each image, means that users will rarely have encountered any taboo-words and the images are therefore not likely to have run through many iterations before the target was

reached. The relative low sample doesn't allow us to draw any certain conclusion, but does indicate that further exploration of similar games is an avenue worth exploring.

It should also be noted that cataloguing practice can vary from institution to institution, and the Sven Türck collection only consists of a single type of staged black and white photography. Other institutions might have formalized policies, e.g. emphasizing narrative descriptions, and a more heterogeneous sample of images might also have yielded different results.

In this paper, the professional descriptors were used as a gold standard set, but further research into the quality of the game-generated tags could entail comparative assessment by end-users between the two types of labels to determine if the non-overlapping terms differ in terms of perceived relevance. The study is indicative of how closely the tags generated by an Output-agreement game resemble professional descriptors and the overall findings suggests that a game like Make a Difference could potentially supplement or perhaps even replace part of the in-house indexing done at cultural heritage institutions with image collections in need of descriptive metadata.

Acknowledgements. This work would not have been possible without help from Jakob Moesgaard, Jacob Larsen and Tom Juul Andersen (The Royal Library in Denmark), funding by the EC (Decision No 1298/2008/EC), helpful advice from Peter Ingwersen (Royal School of Library and Information Science in Copenhagen and Oslo University College) and technical assistance by Jennifer Lea Whisler (Fulbright Fellow at University of Waikato, Hamilton).

References

1. DigitalNZ, http://www.digitalnz.org/make-it-digital/getting-started-with-digitisation
2. Poole, N.: The Cost of Digitising Europe's Cultural Heritage: A Report for the Comité des Sages of the European Commission. The Collections Trust (2010)
3. Driggers, P., Dumas, E.: Managing Library Volunteers: A Practical Toolkit. ALA, Chicago (2012)
4. Brown, P., Hidderley, R., Griffin, H., Rollason, S.: The democratic indexing of images. New Review of Hypermedia and Multimedia: Applications and Research 2(1), 107–120 (1996)
5. Howe, J.: The Rise of Crowdsourcing. Wired Magazine 14(6), 1–4 (2006)
6. Yakel, E.: Balancing archival authority with encouraging authentic voices to engage with records. In: Theimer, K. (ed.) A Different Kind of Web: New Connections between Archives and Our Users. Society of American Archivists, Chicago (2011)
7. Lascarides, M.: Next-Gen Library Redesign. Facet, London (2012)
8. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
9. Enser, P.: Visual Image Retrieval. Annual Review of Information Science and Technology (ARIST) 42, 3–42 (2007)
10. Lancaster, F.W.: Indexing and abstracting in theory and practice, 3rd edn. Facet, London (2003)
11. Estellés-Arolas, E., González-Ladrón-de-Guevara, F.: Towards an integrated crowdsourcing definition. Journal of Information Science 38(2), 189–200 (2012)

12. Oomen, J., Aroyo, L.: Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges. In: 5th International Conference on Communities and Technologies, Brisbane, Australia (2011)
13. Holley, R.: Crowdsourcing: How and Why Should Libraries Do It? D-Lib Magazine 16(3/4) (2010), http://www.dlib.org/dlib/march10/holley/03holley.html
14. Lubna, A.S., Campbell, J.: Crowdsourcing motivations in a not-for-profit GLAM context: the Australian newspaters digitisation program. In: ACIS 2012 : Location, Location, Location: Proceedings of the 23rd Australasian Conference on Information Systems 2012. ACIS, Geelong (2012)
15. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, pp. 319–326. ACM, Vienna (2004)
16. Kazai, G., Kamps, J., Milic-Frayling, N.: An analysis of human factors and label accuracy in crowdsourcing relevance judgments. Information Retrieval 16, 138–178 (2013)
17. Law, E., von Ahn, L.: Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games. In: CHI 2009 Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1197–1206. ACM, Boston (2009)
18. Yuen, M., Chen, L., King, I.: A survey of human computation systems. In: International Conference on Computational Science and Engineering, Vancouver, Canada, pp. 723–728 (2009)
19. Andersen, T.J.: Hvad forestiller billedet? Nyt FB-spil fra Det Kongelige Bibliotek. In: Harddisken, P., Nissen, A.H. (Interviewer), November 23 (2010)
20. Voorbij, H.: Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences. Journal of Documentation 54(4), 466–476 (1998)
21. Kipp, M.: Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords. Canadian Journal of Information and Library Science 29(4), 419–436 (2005)
22. Kipp, M.: Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. Knowledge Organization 38 (2011)
23. Voorbij, H.: The value of LibraryThing tags for academic libraries. Online Information Review 36(2), 196–217 (2012)
24. Iyer, H., Bungo, L.: An examination of semantic relationships between professionally assigned metadata and user-generated tags for popular literature in complementary and alternative medicine. Information Research 16(33) (2011)
25. Lykke, M., Høj, A., Madsen, L., Golub, K., Tudhope, D.: Tagging behaviour with support from controlled vocabulary. In: Proceedings of the 2nd Biennial Conference (Facets of Knowledge Proceedings) 2nd biennial Conference of the British Chapter of the International Society for Knowledge Organization, pp. 41–50. Emerald Group Publishing Limited, London (2012)
26. Thomas, M., Caudle, D., Schmitz, C.: To tag or not to tag? Library Hi Tech 27(3), 411–434 (2009)
27. Wetterström, M.: The complementarity of tags and LCSH – a tagging experiment and investigation into added value in a New Zealand library context. The New Zealand Library and Information Management Journal 50(4), 296–310 (2008)
28. Panofsky, E.: Meaning in the visual arts. Penguin, London (1970)
29. Rasmussen, E.: Indexing images. Annual Review of Information Science and Technology (ARIST) 32, 169–196 (1997)
30. Shatford, S.: Analyzing the subject of a picture: a theoretical approach. Cataloging & Classification Quarterly 6(3), 39–62 (1986)