Rasmus Thøgersen

Crowdsourcing for image metadata; a comparison between game-generated tags and professional descriptors

Supervisor: Peter Ingwersen

Abstract

One way to address the challenge of creating metadata for digitized image collections is to rely on user-created index terms, typically by harvesting tags from the collaborative information services known as folksonomies or by allowing the users to tag directly in the catalog. An alternative method, only recently applied in cultural heritage institutions, is Human Computation Games, a crowdsourcing tool that relies on user-agreement to create valid tags.

This study contributes to the research by investigating tags (at various degrees of validation) generated by a Human Computation Game and comparing them to descriptors assigned to the same images by professional indexers. The analysis is done by classifying tags and descriptors by term-category, as well as by measuring overlap on both syntactic (matching on terms) and semantic (matching on meaning) level between the tags and the descriptors.

The findings shows that validated tags tend to describe 'artifacts/objects' and that game-generated tags typically will 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 higher frequency of 'subjective/narrative' tags, but also more errors.

It was determined that the exact (character-for-character) overlap i.e. the number of common terms compared to the entire pool of tags and descriptors was slightly less than 5% for all types of tags. By extending the analysis to include fuzzy (word-stem) matching, the overlap more than doubled.

The semantic overlap was established with thesaurus relations between a sample of tags and descriptors and adapting this - more inclusive - view of overlap resulted in an increase in percentage of tags that were matched to descriptors. 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 between descriptors and valid tags were either 'same' or 'equivalent' and roughly 20% were associative and 20% were hierarchical. For the hierarchical relations it was found that tags typically describe images at a less specific level than descriptors.

Acknowledgements

A big thank you to my family back home for support.

To the European Commission for sponsoring.

To all the visiting professors from around the globe for answering my questions.

To the good people at KB for providing the great data I got to work with in this thesis.

To all the teachers, administrators, coordinators and facilitators who made the DILL possible and to my co-students in DILL4, who made it completely unforgettable.

To my advisor Peter Ingwersen for inspiring input.

To all the people I have met these last two years.

To Muharrem for Turkey.

To Enrico for sharing the office.

To Hugo for being much smarter than me.

And to Jenny for saying yes.

Table of contents

List of figures

List of tables

Chapter 1

1.1 Introduction

The term crowdsourcing was coined in June 2006 by Jeff Howe in his seminal paper "The Rise of Crowdsourcing" (Howe, 2006), which was published in the trendsetting US tech-magazine 'Wired' - also known for publishing the first article on "The Long Tail" (Anderson, 2004), another popular web 2.0 neologism.

The original crowdsourcing piece mainly focuses on business perspective, giving an example of how a company can reduce costs dramatically by outsourcing certain processes to the crowd, rather than having highly trained (and thus costly) professionals perform somewhat menial tasks. This emphasis on the monetary aspect is hardly surprising given the nature of the magazine and how the precursor of the word – outsourcing – has heavy business connotations. Howe does however mention examples from the non-profit sphere. One of these being Wikipedia, which epitomizes crowdsourcing and today, seems almost synonymous with the concept.

My early personal experience with Wikipedia, aside from increasingly using it as the go-to source for quick reference, was during my undergraduate studies at the Royal School of Library and Information Science in Copenhagen, Denmark. In an introductory course on Information Retrieval, which was my first introduction/exposure to many of the concepts which now permeate most of my professional life, we spent many sessions discussing the impact that the World Wide Web in general and Google and Wikipedia in particular, had had on our profession. At the time, I was working reference at a public library and found the shortcuts offered by the sleek search engine and the vast encyclopedia quite useful. My professor, however, did not agree. Her beliefs were firmly planted in another paradigm, where library systems and the objects they contained were hard-to-access 'things', requiring a professional intermediary to interact with them. We spent a substantial amount of time doing searches in Thompsons Dialog (the classic command-driven, text-only interface) and discussing the importance (and widespread lack) of information literacy in higher education, which made librarians completely invaluable as gatekeepers of quality information.

At one time my professor even appeared on Danish National Television in primetime to warn the general public against the inherent bias, the lack of validation, the mistakes, the exaggerations and the outrights lies, found on and in between the lines of the massive sprawling mess that constitutes Wikipedia. This was around the same time as the publications of the now infamous article in Nature, which compared the electronic fledgling to the old paper giant of Encyclopedia Britannica, finding an equal amount of factual errors in their articles (Giles, 2005). While we never saw eye-to-eye on the emerging technologies, I should note that I did take a lot with me from that course that I still rely on today when dealing with more complicated tasks and that it, in many ways, formed my way of thinking about information retrieval.

One year later I found myself on an exchange semester at the School of Information and Library Science at the University of North Carolina, Chapel Hill. Due to problems with scheduling I once again found myself taking an introduction level course on Information Retrieval. As my English at the time was shaky and the other courses I had enrolled for at a higher level, I decided to follow the course, if for nothing else, to observe differences in didactic and pedagogical approaches to what would presumably be a somewhat similar curriculum to the one I had delved into one year prior. In terms of

actual IR I didn't learn much in that course – as the core concepts proved to be very similar - but this was outweighed by the radically different approach to the library profession. Our entire course was organized in a wiki; our homework was writing blog-entries and commenting on our fellow students' writings; and the final assignment was actually to identify a concept from the broad LIS-field, not yet described in Wikipedia, and then to write the missing article in the real world Wikipedia. I ended up writing an article on Functional Requirement for Subject Authority *Records*, which at the time was known as FRSAR, but has since then been changed to Functional Requirement for Subject Authority *Data* (FRSAD). In the following weeks after submitting the article to Wikipedia and the URL[1] to my professor for evaluation, I remember following its climb up the Google-ranks until it finally made it into the first page of hits. Other contributors have later updated my original article and our piece consistently[2] ranks in the top five hits on Google for both FRSAR and FRSAD. This tiny accomplishment whetted my appetite for crowdsourcing, not only as a participant, but also as a subject for study.

My time in North Carolina was in general all about embracing and utilizing and it was saturated with an unbridled enthusiasm for what librarians can achieve if they collaborate with, rather than fight against, the tide of emerging technologies. To stay in the realm of ocean related metaphors, we can rewrite the famous John Dunne quote: "*no library is an island*", nor should it be.

As an international master student in (as well as avid observer of the development of) digital libraries I have noticed how many librarians now fully embrace the users of their collections as more than passive recipients of information. In the fall of 2011, Wikipedia toted the banner "Wikipedia loves libraries," marketing an event spanning all of North America in which the public was invited into libraries to participate in so called editathons, resulting in the creation of a large quantity of Wikipedia articles based on information found in library collections. The idea was based on an event at the British Library in June 2011 which indicates that, at least in the English speaking parts of the international library community, there is a growing interest in and acceptance of crowdsourcing.

For me personally, this feeling of acceptance culminated during the 2011 Europeana Tech Conference at the Austrian National Library in Vienna, where I participated in an entire session dedicated solely to Distributed Community Empowerment (i.e. crowdsourcing) as a method, displaying the myriad of different ways in which cultural heritage institutions across Europe attempt to apply it to their context.

Back in Denmark, at the Danish Royal Library (KB), this also held true. Inspired by a lecture by crowdsourcing pioneer Luis von Ahn, the decision was made to create a Human Computation Game in the same vein as von Ahns ESP-game: an innovative online game allowing for easy and fun tagging of large quantities of images. The Deputy Director General and head of Information Technology Services Birte Christensen-Dalgaard wrote an article in the Danish Journal for Research Libraries 'Revy' about crowdsourcing and how libraries could embrace this new approach (Christensen-Dalsgaard, 2010). Shortly thereafter, the game 'Gør en forskel'[3] was launched to the general public via Facebook. Serving the dual purpose of both gathering money for Red Cross and annotating the

[1] http://en.wikipedia.org/wiki/FRSAD
[2] Yes, I still check once in a while
[3] Danish for 'Make a Difference'

cultural heritage (Andersen, 2010), it ran for about two weeks and resulted in more than 22.000 tags being added to approximately 2000 recently digitized images.

This thesis is about the output of that game. It is also about metadata, folksonomies, image indexing, validated tags, descriptors, human computation games and, of course, crowdsourcing.

1.2 Digitization of cultural heritage in Denmark

The digitization effort in the Danish GLAM sector began around 1990 with digitization of printed registries and catalogs. It was originally a way to make collection metadata machine-readable, and thereby making internal tasks such as collection management and retrieval more efficient. The advent of the WWW and its widespread adoption from the mid-nineties and onward, added a new layer of functionality to the digitized metadata and helped make the collections accessible to a wider audience, culminating in the launch of the Danish union catalog Bibliotek.dk in October 2000, which allowed for fast nation-wide interlibrary loan service via the Z39.50 protocol.

At the same time as the developments in the exchange and usage of metadata, Danish cultural institutions starting scanning the actual objects - rather than just the catalogs and records describing them. One of the first examples of this was Den National Billedbase, which was established in 1993. It tapped into the collection housed at Kort- og Billedafdelingen at KB - home to more than 18 million maps, photos, paintings and other kinds of illustrations representing a vast collection of ideas from worldviews collected throughout the latter part of the last millennium. By September 2011, almost 100.000 images were scanned and stored at Den National Billedbase.

Many of the early digitization efforts were mainly done with preservation in mind – the objective was simply to ensure the continued survival of the cultural heritage for posterity; digital copies stored first on magnetic tapes, optical storage unit or servers, were deemed to be less vulnerable than physical copies. This initial focus on preservation can be attributed to the fact that electronic access still was in its infancy.

Like the shift from a strictly internal functionality to a combination of internal and external functionality was fueled by an increasingly IT-oriented user-group, another shift occurred relating to the digital objects themselves. While preservation still is seen as a priority, dissemination and access have been increasingly recognized as important focal areas, as seen in the 2006 media policy agreement:

"Extending on the 2003 assessment on preservation of the cultural heritage, a composite task force – managed by the Ministry of Culture – with representatives of the relevant public authorities, institutions and organizations will be tasked with preparing different proposals for digitizing selected, prioritized parts of the cultural heritage in order to preserve, disseminate and provide access to them."[4]

The composite task force mentioned above published their final report on digitization of the cultural heritage in 2009, in which they recommended a focus on hard-to-reach materials, not otherwise accessible to the general public. As the majority of books in KBs national library collection are catalogued and are obtainable via the aforementioned Danish union catalog in a matter of days,

[4] Translated from Danish

focus should be mainly on fragile materials not suited for transportation. The selection criteria, or what could be described as a sort of informal collection development policy for digital cultural heritage, are generally geared toward making cultural heritage accessible.

The media policy agreement lists a number of technical necessities, such as high quality scanning and safe long-term storage, but also mentions exposure and visibility in popular search engines (e.g. Google) and cultural portals (e.g. Europeana) via collected metadata. This again underlines the idea of opening up the library to a wider context, as mentioned in the introduction.

The dual goals of dissemination and access are reflected in the 2008 decision at KB to establish a new image database with added metadata to improve browsing, discovery and display of the digitized images. The new, improved service is currently housing approximately 40.000 images, predominantly from older collections, as the publication of out-of-copyright materials is a lot less complicated from a legal point of view.

One of these collections, of which a subset recently was scanned and indexed, is the work of the Danish photographer Sven Türck.

1.3 The Sven Türck collection

The current head of Kort- og Billedafdelingen Mette Kia Krabbe Meyer wrote her PhD on various aspects of photography in Denmark in 1910-1950 and described Sven Türck as 'one of most prominent exponents' of the fledgling discipline advertisement-photography in the 30's and 40's (Krabbe, 2004, p.7). Türck used the mindset from the advertisement industry when it came to his documentation of ordinary Danish life, portraying (selling) an idyllic version of Danish life. He used set-pieces to show healthy looking and smiling representatives of the working class, families, gymnasts and youth, as well as summer landscapes and national symbols such as the Little Mermaid, Tivoli and the stork. His pictures are continually sought after as illustrations for historic publications and their copyright status makes them suitable for access-oriented digitization.

As of March 2012, 4177 of the Sven Türck photographs have been scanned; described thoroughly by professional intermediaries; and been made available via the new image database, as seen in figure 1.

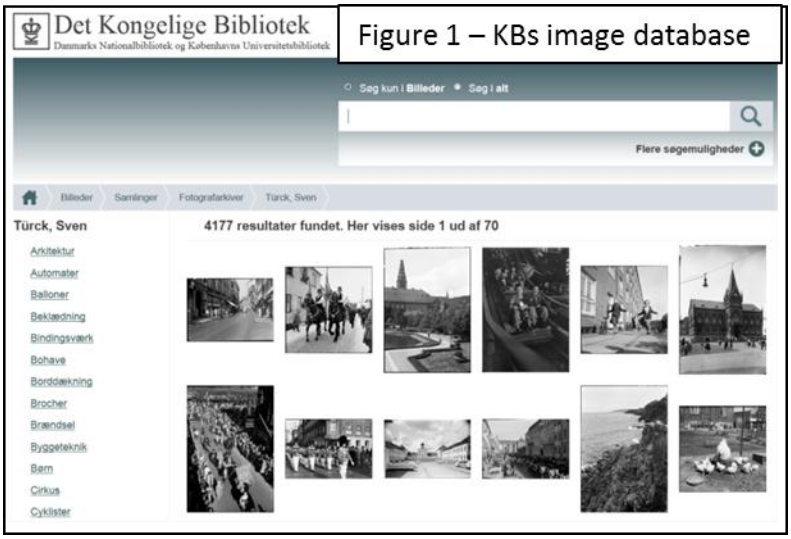

Figure 1 - KBs image database

Aside from the Google-like search box dominating the top of the screen, the left side of the interface is composed of an alphabetized subject hierarchy of descriptors intended for browsing up and down different levels. The descriptor "beklædning" (clothing) can for instance be unfolded to reveal 728 images with various kinds of

clothing described. Some of these are even more specific, such as "bukser" (pants), which is divided into six kinds of pants.

Each image has a number of descriptors assigned to it, in addition to other bibliographic information such as title, author, year, notes, location and person portrayed (mainly if the image is portraying royalty). This allows for browsing as well as keyword, and phrase-searching within the metadata. The descriptors and their underlying hierarchical structure is an ad hoc endeavor, suited for this specific collection and does not stem from any controlled vocabulary, but it does represent the viewpoint of professional indexers at KB.

The Sven Türck collection happened to be in the technical pipeline when GeF was launched, so it was chosen as the collection to be annotated by the game (Andersen, 2010).

1.4 Gør en Forskel (GeF)

GeF was developed by KB as a way to raise money for charity and annotate the cultural heritage at the same time (Andersen, 2010). It was inspired by the ESP-game and follows the same simple gameplay[5]:

1. You and a partner see the same image
2. Each of you must guess which word your partner is typing

You are assigned points for every time you 'guess' a word, meaning you both work together towards a common goal. While the ESP-game relied on synchronicity, GeF allowed for players to assign tags independently of each other, rendering a tag validated when three players, independently of each other, had used it to describe a given picture.

Aside from creating an incentive to tag by making it a game, the point of the game is this idea of the validated tags. Tags - and the folksonomies they form - are uncontrolled in nature and thus represent a stark contrast to the traditional controlled vocabularies used in libraries, which place strong emphasis on underlying structure, non-bias, consistency and correctness.

If two strangers, independently of each other, confirm a tag, this should eliminate some of the undesired idiosyncrasies of the free tags (i.e. misspelled, personal, non-sense and sabotage tags) and make them more 'suitable' for serving as access points in the library OPAC.

Figure 2 displays an example of three different players' tags which results in a single validated tag 'Polar bear' (isbjørn). As seen, the chosen validation threshold was three (i.e. three players have to agree on a term before it is considered valid. This validation threshold is an arbitrary number x between 1 and n, which determines

Figure 2 – Example of validation process

[5] From http://www.gwap.com/gwap/gamesPreview/espgame/

the number of pairs that have to agree on a tag/label, before it is considered valid. If x=1, then one pair has to agree and the threshold is very lenient. If x=40, a word has to be repeated significantly more times before it is considered valid (von Ahn, 2006). Figure 2 show the single validated tag, but it also shows how the term 'Fur' occurs twice between the three players and how they added a total of 12 completely free tags. These three tag-types i.e. tags at various degrees of validation (from now on referred to as Free tags, 2Vtags and 3Vtags) together with the professionally assigned descriptors, are the

A private foundation had donated 5000 DKK, with each validated tag (in this case meaning 3Vtags) resulting in a 2 DKK donation to Red Cross, putting a ceiling of 2500 validated tag on the project; once 2500 tags had been verified at the given threshold of three, the 5000DKK would be donated and the game would be shut down.

Prior to the 2010 launch, an initial run was completed internally at KB in December 2009 to serve as a beta-test of the 'real-world' launch the next year. After opening the game to the public, KB had to close it after a week, as the quota for validated tags was filled. In the timespan of the game, the 2079 images were tagged 22787 times with free tags. When KB shut down GeF, 2516 tags had been validated at the set threshold (Moltved, 2011).

In terms of speed and the sheer numbers of tags generated – as well as from Red Cross' point of view – the game can be said to have been very successful.

1.5 Justification of research

The output, the verified tags, seem to represent a compromise between the controlled descriptors found in catalogs and free tags found in popular collaborative information services or folksonomies such as LibraryThing, Flickr or Citeulike[6].

As will be discussed in more depth in chapter 2, a substantial amount of scientific effort has been dedicated to research into the nature of these folksonomies in recent years. This research has according to Veres (2006) ranged from "*mathematical approaches for clustering, and identifying affinities, social theories about the cultural factors in tagging, and cognitive theories about their mental underpinnings*"(p. 325).

A recurring theme within Library and Information Science (LIS) is on identifying affinities and discovering how tags/folksonomies relates to descriptors/ontologies, which is a very logical approach considering the practical implications of these relations. If a library can harvest high quality tags from folksonomies, they can either avoid costly indexing or enrich their existing catalog records (Steele, 2009), which is why researchers have investigated overlap and relationships between LibraryThing and LCSH (Lu, Park, & Hu, 2010), Delicious and LCSH (Yi & Chan, 2009); tags generated from 2.0 enabled OPACs and LCSH (Thomas, Caudle, & Schmitz, 2009); Flickr tags and indexer assigned descriptors (Rorissa, 2010); and MeSH terms and CiteULike Social Tags (Lee & Schleyer, 2010) to name a few.

Human Computation Games have mainly been discussed in the discipline from which they originate, computer science, and the original purpose of these games was actually to provide data to improve machine learning algorithms (von Ahn & Dabbish, 2004).

[6] These examples of folksonomies will be explained more in depth in chapter 2.4.2

To the best of my knowledge, no studies have been undertaken to discover the relation between the outcome of a Human Computation Game and the outcome of professional indexing[7]. This can be attributed partly to the fact, that these games are rarely applied in situations where professionally created metadata already exists. The other possible reason is that, unlike the studies already mentioned on folksonomies – that in their nature are open and typically allow for harvesting of data via APIs – while the data generated by GWAP is not freely available to everyone.

In this aspect, the Sven Türck collection, having undergone indexing from both professional indexers and players of the game, is unique. The reason for choosing Sven Türck's photographs instead of an un-indexed collection for the game was purely technical – it was in the pipeline (Andersen, 2010). However, this coincidence is fortunate from a researcher's standpoint, as the resulting metadata represents a very interesting and unique insight into the relation between a very specific kind of tags and the descriptors assigned by professional indexers.

In a study on tagging in museums, Trant (2006) articulates why cultural heritage institutions should study their own indexing viewed in the light of the user-generated kind, writing that: "*Looking at the types of tags supplied by those outside museums and studying how they correlate (or do not) with data now made available by museums can provide insight into users' perceptions, identify areas of disconnect, and help museums adapt to meet their missions*" (p.86) and an even more direct purpose is given by Wetterström (2008):"*User-assigned tags could provide additional access points, and the co-existence of tags and controlled vocabularies... could thus enhance the discovery of documents "*(p.297).

The act of actively pursuing metadata by developing specialized software, rather than passively harvesting it from external platforms (e.g. via folksonomies APIs) or from next generation OPACs (e.g. by allowing user-rating, commenting and tagging of objects within the catalog itself), is an example of how libraries tread new ground by deploying crowdsourcing techniques.

A tagging game is one possible way to enrich our digital cultural heritage by crowdsourcing, but in order to evaluate the efficiency/potential of such a game, we must deepen our understanding of the output in relation to professional indexing, as well as try to understand how one of the key components 'the validation threshold' affects the tags.

This is done by looking not only at the validated tags at the predetermined threshold in GeF (3Vtags), but also by including all tags (Free tags) as well as tags at a more lenient threshold of two (2Vtags) in the investigation, and then considering the three kinds of outputs as separate sets of data, eavh to be compared with a fourth kind - the existing metadata. Descriptors already attached to the Sven Türck collection by professional indexers at KB:

[7] The closest example is the Dutch crowdsourcing game "Waisda?" which was a similar type of game created for video content from the Netherlands Institute for Sound and Vision. The researchers evaluated the verified tags from "Waisda?" with the help of a professional senior cataloguer, who assessed the tags (Gligorov, et al., 2010, s. 5).

Figure 3 – The four different datasets chosen for analysis

Thus the purpose of this thesis is to determine how similar the different image metadata output of a Human Computation Game (tags at three different validation thresholds) is to the output of professional indexing (descriptors). I will do so by investigating on both vocabulary and object level, by asking the following questions:

1.6 Research questions

RQ1: To what extent do tags (across all three validation thresholds) and descriptors fall within the same term-categories?

The term-categories (see chapter 3.3.1.1) are derived from the literature review of similar studies regarding classification of image index descriptors and modified during analysis to fit the data. Answering this question can uncover differences in the vernacular vocabularies used by players and indexers, as well as provide information on differences between different levels of validation.

Once the distribution of tags and descriptors among term-categories have been determined, which is an effort done on vocabulary level, focus will turn to object level and first determine one kind of overlap between players and indexers, by asking:

RQ2: To what extent do players and indexers use the same terms to describe the same image?

This is done by determining the 'syntactic exact and fuzzy overlap' between tags and descriptors assigned to each image. Exact overlap refers to when tag and descriptor match each other character for character. The fuzzy overlap is measured to determine if tag and descriptor share the same word-stem. Both exact and fuzzy matching is done for each picture and compares Free tags with descriptors, 2Vtags with descriptors and 3Vtags with descriptors. Having determined the syntactic overlap, the final research question addresses how similar the metadata output is on the semantic level i.e. whether they express similar concepts:

RQ3: To what extent do players and indexers use thesaurus-related terms to describe the same image?

This serves as a natural extension of RQ2; by first seeing whether player and indexer use the exact same term when describing an image, and then widening the scope to subsume different forms of the same terms, RQ3 go one step further and asks how the similarity is affected by looking not just at the terms, but at underlying meaning behind them and determines whether there is a higher correspondence. This final question revolves around 'semantic overlap' between players and indexers.

1.7 Limitations

This is an initial foray into the relationship between two kinds of metadata.

It must be stressed that measuring the quality or value of tags and descriptors is not part of this research. This thesis does not compare the quality of the tags and descriptors, as this would have to be done by user-evaluation, which was impossible due to time constraints.

Even though some degree of image analysis was needed to determine whether belonged to the term-category 'From image' term-category (see chapter 3.3.1.1), the majority of the analysis was done purely amongst terms, effectively omitting the described objects themselves, meaning that the images themselves were rarely studied.

Other studies using similar methods (Lykke, Høj, Madsen, Golub, & Tudhope, 2012; Rorissa, 2010) usually had more than one researcher and relied on dialog and feedback to agree on term-categories and thesaurus relations, something not possible within the limited scope of this thesis.

The Sven Türck collection itself also represents a very specific type of images: Danish photography from a certain time period, all in black and white. Similar studies done on different or more heterogeneous collections might yield different results.

1.8 Outline of thesis

The first chapter of this thesis provides a rationale for the study through background information, and context to the work as a whole. The research problem, the objectives and research questions of the study are stated and the perceived limitations are described as well.

Chapter 2 reviews the literature that is pertinent to the topic and that has informed this study.

The third chapter outlines the methodology used in this research project. Both the data collection and the research design are accounted for.

Chapter 4 comprises the findings and discussion of these.

The final chapter presents conclusions from this research project and offers suggestions for future research topics derived from the thesis, as well as recommendations for usage of the GeF-output.

Chapter 2 – Literature Review

The introduction to this thesis ended with the following statement: "*this thesis is about the output of the GeF Game. It is also about metadata, indexing, folksonomies, image indexing, validated tags, human computation games, tags, and, of course, crowdsourcing*."

This sentence will form the backbone of my literature review, as I will try to cover these different topics to the extent it is needed to frame the research at hand. I will start by describing the overarching theme, indexing, from the two relevant perspectives, namely traditional indexing and folksonomies and then turn to the specific challenges when it comes to image indexing. Finally, I will describe crowdsourcing and the Human Computation Games.

2.1 Metadata

Metadata or data-about-data can be expressed in various ways/schemas, depending on the community (Intner, Lazinger, & Weihs, 2006, ss. 21-61) and numerous formats (txt, rdf, xml, html) depending on the system(s) using the record. In the library community, MARC or a local variation of MARC is the 'lingua franca' (Matthews, 2000, s. 19), but the evolution into digital libraries has led to the adaptation of other schemas. Common examples include Dublin Core - which due to its simple nature allows for easier interoperability and exchange across institutions/collections – and METS, which is a container format able to hold other schemas within it.

A traditional distinction is made between two roles in creation of metadata: the cataloger and the indexer. Whereas cataloging is about taking explicit information - such as author name or publication date - from a given source (e.g. from the title page), indexing is about extracting implicit information regarding the subject, theme or topic of the object. In practicality, the terms tend to get mixed up, and the tendency to use the term "indexing" for any bibliographic access technique, coupled with the practice of calling index terms "descriptors" does not help in providing clarity (Olson, Boll, & Aluri, 2001, s. 125). The cataloger and indexer can also very well be the same person and the entire process can occur at one time. In other cases, the subjects (and descriptors) are determined by an expert, while the explicit information is written by a cataloger.

A usual division of categories in digital metadata is between: descriptive, structural and administrative (Tennant, 1998). The structural refers to how an object is structured e.g. a digitized book, where each page is a separate file. The administrative metadata is used in managing the object and includes information about acquisition and legal status. The descriptive metadata resembles the descriptive MARC format and is where the information added by catalogers and indexers is found, i.e. bibliographic data and descriptors.

While the metadata landscape is changing with technological advances, the purpose of descriptive metadata remains the same: to act as a surrogate record which can be used to retrieve an object in a collection.

Below is such a record, taken from Library of Congress:

```
000     01164cam a22002654a 450
001     16280935
005     20111020130319.0
008     100611s2011 mauad b 000 0 eng
906     __ |a 7 |b cbc |c orignew |d 1 |e ecip |f 20 |g y-gencatlg
925     0_ |a acquire |b 2 shelf copies |x policy default
955     __ |b rg11 2010-06-11 telework |i rg11 2010-06-11 telework to Dewey |w rb07 2010-06-14 |d xd pbk
ISBN delete 2010-08-18 |d pc11 2010-09-27 index removed |a xe08 2011-06-07 1 copy rec'd., to CIP ver. |f CIP
ver. re03 2011-06-17 To BCCD |t rf06 2011-10-20 to BCCD (Copy 2)
010     __ |a 2010022788
020     __ |a 9780262015097 (alk. paper)
040     __ |a DLC |c DLC |d DLC
050     00 |a Z665 |b .L36 2011
082     00 |a 020.1 |2 22
100     1_ |a Lankes, R. David.
245     14 |a The atlas of new librarianship / |c R. David Lankes.
260     __ |a Cambridge, Mass. : |b MIT Press, |c c2011.
300     __ |a xv, 408 p. : |b ill. (some col.), col. maps ; |c 26 cm. + |e 1 chart (67 x 89 cm., folded to 23 x18 cm.)
504     __ |a Includes bibliographical references.
650     _0 |a Library science |x Philosophy.
650     _0 |a Library science |x Forecasting.
650     _0 |a Libraries and community.
650     _0 |a Libraries and society.
```

Figure 4 – Example record from LOC

The first fields (000-008) are machine-readable values containing information regarding date (005+008) and control numbers for the item (001) and the record itself (000). After that, the local processing field (906), local selection/retention field (925), and local tracking field (955) provide administrative metadata regarding acquisition and provenance. The 010 field is a control number relating to authority control, and in the 040 field, we find information about the creator of the record itself (in this case LC=Library of Congress).

All the information so far is fairly straightforward and can be done without actually having a copy of the book in hand. For the remainder of the record I have highlighted the information noted by a cataloger in red and by an indexer in green.

The cataloger has to follow strict guidelines (in this case AACR2 or RDA) when lifting the explicit information from the book and has to actually hold the book and open it to see if it contains bibliographic references and illustrations.

The 650 fields, marked in green, contain the descriptors. In 050 and 082 we see classifiers, numerical representations of subjects. It is important to note that an indexer – in theory - could assign any kind of sign to an object. If the color red or the sound of a blackbird singing was found to express the subject of the book better (and the GUI of the catalog could support retrieval of those signs) that would be indexing too.

This example is provided to show the end-goal of the indexing process. Indexing starts with an object and it ends with a number of index terms, describing what the object is about. In this thesis, I am considering indexing as "*the process by which the content of an information resource is analyzed, and the aboutness of that item is determined and expressed in a concise manner*" (Taylor & Joudrey, 2009, s. 22).

2.2 Indexing

Surrogate records are no longer found solely in library catalogs and indexing is no longer just done by catalogers and indexers.

Ingwersen (2002) classifies the kind of indexing done at LOC as 'interpretation of content with a purpose'. In addition to the interpretation by the indexer, he adds three more 'types of aboutness'; author aboutness (the content 'as it is' e.g. the kind of aboutness used in automatic indexing), request aboutness (content expressed by the query that finds it) and finally user aboutness i.e. the users interpretation of the content (p .289). This last one is the aboutness expressed in a folksonomy.

The same is illustrated by (Stock, 2007) who illustrates the different ways in which a document can be 'seen'.

Figure 5 is a modified version of the model found in Stocks article, with the addition of a fourth leg for the 'Query/request' based indexing.

The Content Based Image retrieval Is the automatic extraction of

The difference between the left (folksonomy) and the right (ontology) side is what is most relevant for this thesis and these two fundamentally different indexing strategies will be explained in the subsequent sections.

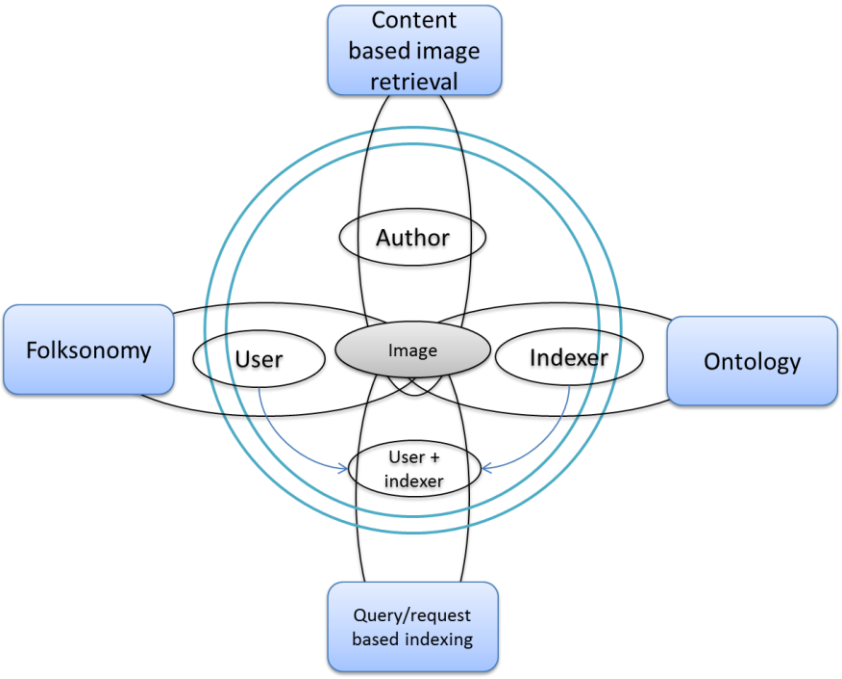

Figure 5 – Four perspectives on indexing. Modified from Stock (2007).

2.3 Traditional indexing

According to Lancaster (2003, p. 9) traditional subject indexing is a two-step process:

1: Conceptual analysis
2: Translation

These steps may not be that explicit and may occur at the same time, but intellectually they are different from one another. The conceptual analysis is about deciding the 'aboutness' of the object at hand. This process of determining the subject is arguably more art than science. Each part of the object (title, index, the object itself) or even external sources such as reviews can provide hints as to what the true subject matter is. It is up to the indexer to "…t*ake these*

hints, treat them in a systematic manner, and come up with an effective assessment" (Intner, Lazinger, & Weihs, 2006, p. 110) to identify the appropriate subject. This systematic manner also involves a second step, which is the translation of the results of the conceptual analysis into index terms, usually taken from some sort of controlled vocabulary e.g. a thesaurus.

These two steps can be broken down even further. David, Giroux, Bertrand-Gastaldy, Lanteigne, & Bertrand (1995) sheds light on the process of indexing by approaching it as problem solving activity in multiple steps within a knowledge space, each requiring a different type of expert knowledge:

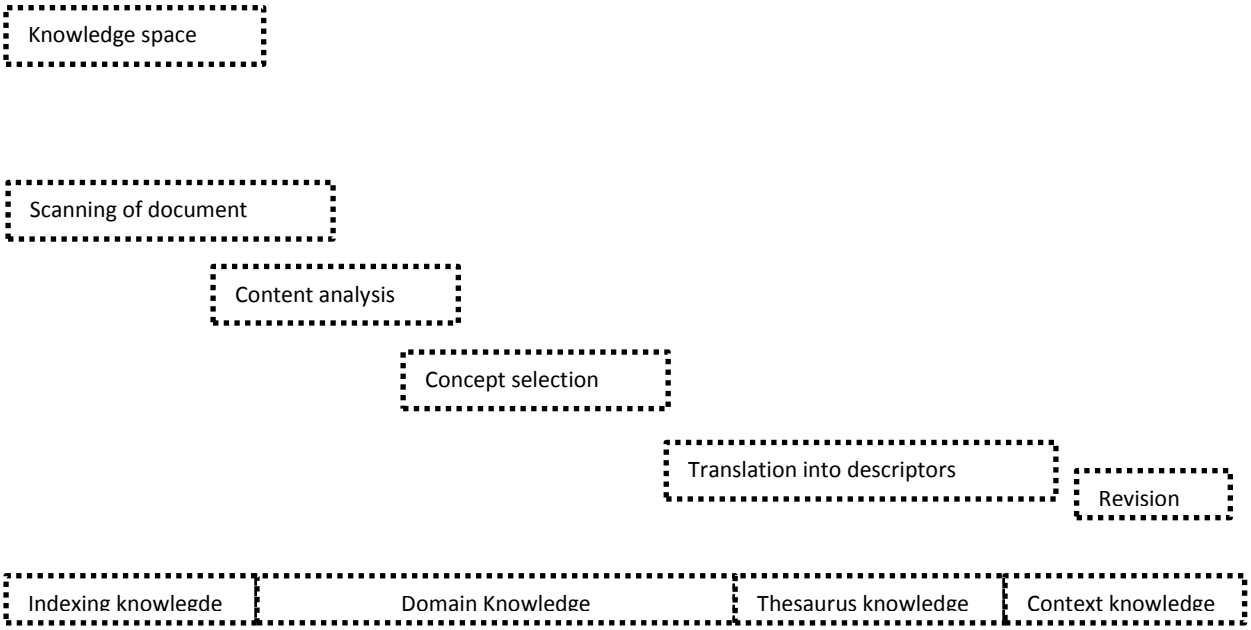

Figure 6 – Indexing process (David, Giroux, Bertrand-Gastaldy, Lanteigne, & Bertrand 1995).

Their model shows "*examples of the form of knowledge used by an expert indexer familiar with a particular domain, a particular thesaurus and a given working environment*." The operational term here is 'expert'. The entire knowledge space in which indexing takes place is permeated by the notion of specialist insight - i.e. indexing is not for the uninitiated. This stands as a contrast to the democratic approach taken by folksonomies, which will be discussed in subsequent sections of this review. It should also be noted, that while the model is clear, real life indexing might be a lot more fuzzy, as the actual process is less step-by-step and more iterative e.g. can domain knowledge already be applied while scanning the document and can context knowledge be used when concepts are selected.

An example of the context knowledge is the chosen indexing policy for the collection (which can either be explicitly written down in a guidebook or simply taught by neighbor training). The indexing policy often dictates the exhaustivity and specificity at which the object is to be indexed.

Exhaustivity is the number of descriptors that will be considered in the analysis. A book will rarely be about just one concept, which is why several descriptors combined make for a richer description. This breadth of coverage can be either selective, with only what the indexer considers to be the most important concept, or truly exhaustive, denoting every single concept. Is the book only about "multilateral aid" or is it also about "international relations", "corruption", "agriculture" and "cultural differences"? Specificity, on the other hand, is related to the semantic 'depth' of the indexing, i.e.

how general/specific the assigned term is. These two factors affect the precision and recall in retrieval – precision being a measurement of how many of the objects that a given query returns are relevant, and recall describing how many of the relevant documents in the entire collection the query retrieves. A very exhaustive indexing policy might result in a very high recall, but also lower precision. Conversely, too specific terms might yield very precise results, but lower recall (Svenonius, 2000).

2.3.1 Ontology knowledge

A key element in the indexing model is 'thesaurus knowledge', which can be replaced with 'ontology knowledge' for our purposes; as indexers can rely on other types of knowledge organization schemas than a typical thesaurus. In the Danish National Bibliography for example, the descriptors assigned to objects simply come from a list of 'controlled' words, in the sense that a central authority have acknowledged them. The list of subject terms is an alphabetically sorted list of thousands of words, without any relation other than they appear in the same list.

Thomas Gruber (1993) defines an ontology as an: "*explicit specification of a conceptualization*" (p.199) meaning that it describes concepts and relationships existing within a community. One can say that the conceptualization is an expression of some semantic structure, which encodes implicit knowledge from the community. The list of controlled descriptors used in Denmark is an ontology, albeit a very simple one, as no semantic structure exists, other than the alphabetic sorting and that they belong to the same class[8] of descriptors.

McGuiness (2003) lists the different kinds of ontologies along a range of increasing expressivity.

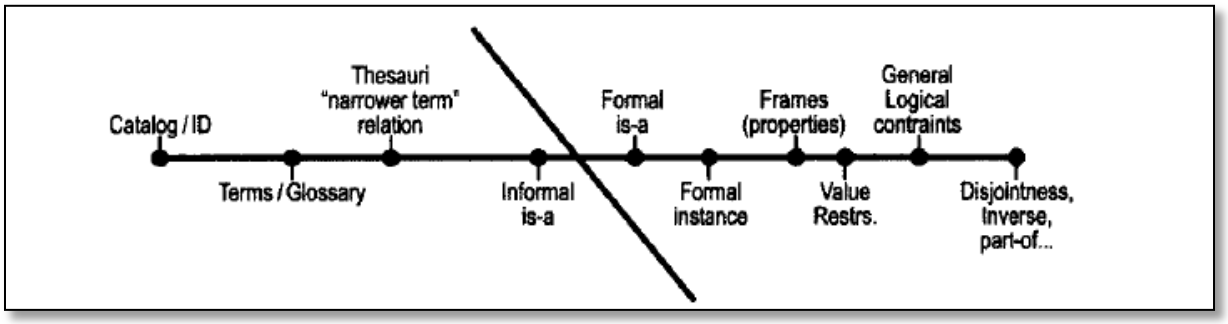

Figure 7 – A taxonomy of ontologies (McGuiness, 2003)

In this taxonomy we find the aforementioned descriptors at the leftmost of the continuum. Had the descriptors been defined somehow, with meaning added to them in natural language (like a dictionary), the list would have been slightly more expressive. Thesauri add sematic relationships to glossaries, often in the form of synonymy between two terms, pointing out how one term should be used rather than another. Thesauri are traditional examples of controlled vocabularies, with a strict hierachical structure (Broader and Narrower Terms) as well as horizontal relationships (Related Term or Use For). They are either contructed based on literary warrant (inherent to the documents in a given collection), user warrant (derived from user information behavior in relation to the collection) or a combination of the two.

The advantage of adding semantic relationships is apparent when we are talking access, as this elaboration allows for query expansion, e.g. "did you mean?", or faceted browsing up and down a

[8] Catalogers can also opt to use uncontrolled keywords.

hierarchical set of descriptors (as shown in the Sven Türck collection in chapter 1.3). As a significant part of my methodology revolves around thesaurus relations, an in-depth explanation of all of them will be given in chapter 3.3.3.

Moving further to the right of the diagonal line in McGuiness' taxonomy we find increasing levels of complexity, in which one, for instance, can start restricting values, which can help machines to infer things about the terms. In a library system one can, for example, state that all personal authors have to belong to the class people.

Each of these types of ontologies can be used for systematic classification by an indexer, and their application is dependent on organizational and technological context. Ontologies can be specialized to either cover a certain type of object, like the Thesaurus for Graphic Materials (TGM); a certain discipline, like Medical Subject Heading (MeSH); or modified for a specific collection or organization. Ontologies can also be universal, in the sense that they attempt to cover any given topic. Examples of these from the library-world are the Dewey Decimal Classification (DDC) system and the Library of Congress Subject Headings (LCSH).

Another well-known example of a comprehensive, universal (English) ontology is the Princeton-developed lexical database WordNet, which resembles a thesaurus in its structure. The current version of WordNet holds 147.278 nouns, verbs, adjectives and adverbs[9] and their relations.

The Danish DanNet does the same for Danish as Wordnet does for English and is a collaborative project, undertaken by the Literary and Linguistic Society under The Danish Ministry of Culture and the research institution Center of Language Technology, University of Copenhagen. It is meant to exploit data from earlier projects, The Danish Dictionary (DDO) and SIMPLE.DK.

Of particular interest to this thesis, is the online visualization of the data found in DanNet: AndreOrd which allows users to access 66.300 terms with 326.652 different kinds of relations. This tool serves as the reference standard for the categorization of thesaurus relationships (see section 3.3.3.2).

2.3.2 Catalog objectives

Regardless of which ontology with which we describe the process of indexing, it starts with an object and ends with descriptors in a record of some sort. In libraries we index (and catalog) our collections in order to facilitate discovery and access to them. Without useful access-points, our shelves and data are out of reach of the users and therefore meaningless.

The act of creating document representation fits into the larger framework of information retrieval activities in general (Lancaster, 2003, p. 5) and serves to fulfill the objectives any catalog must fulfill, as suggested by Charles Cutter in his 1876 *Rules for a Dictionary Catalog* as cited by Taylor & Joudrey (2009), p. 45:

1. To enable a person to find a book of which either
A. the author is known
B. the title is known
C. the subject is known

[9] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

2. To show what the library has
D. by a given author
E. on a given subject
F. in a given kind of literature

3. To assist in the choice of a book
G. as to its edition (bibliographically)
H. as to its character (literary or topical)

While Cutter's objectives have since then been refined (Paris principles and FRBR *insert citation)*, the idea of a *subject* remains as a central point in any effort to describe the content of a library collection. Svenonius underlines the importance of subject as rivaling "…*author in importance in organizing documents and providing access to them. That all documents on the same subject must be displayed together is mandated by the collocation objective*." (Svenonius, 2000, p. 46). One aspect of the metadata does not take precedent over the other, but relates to the type of information search a user is performing.

2.3.3 Descriptors

In the 650 field from the example record we have two instances of the topic "Library Science" with the different subfields "philosophy" and "forecasting". Those are two examples of how LCSH can use subfields to parse subdivisions. Both of these are the 'x' general subdivision and a complete list of possible LCSH subdivisions is seen below:

- $a - Topical term or geographic name entry element
- $b - Topical term following geographic name entry element
- $c - Location of event
- $d - Active dates, or the time period during which an event occurred
- $e - Relator term, which specifies the relationship between the topical heading and the described materials, e.g., depicted
- $v - Form subdivision
- $x - General subdivision
- $y - Chronological subdivision
- $z - Geographic subdivision

This type of division allows for descriptors to have a predisposed meaning, which can serve as a qualifier for the descriptor and make specialized subject field searching possible. Furthermore, descriptors come with a standardized way to spell words and express subjects that facilitates the collocation-on-subject objective, which can only be done if the term expressing the subject is constant.

Like mentioned before, increasingly expressive ontologies allow for more advanced options when designing the systems that utilize the descriptors. A hierarchically organized set of descriptors allows for browsing up and down in the same hierarchy e.g. to identify more specific subjects, and if we move all the way up the scale to modern ontologies, like OWL, the possibilities (and challenges for indexers and system designers) increase dramatically (Nagarajan, Verma, Sheth, & Miller, 2006).

2.3.4 The Semantic gap

Bates (1998) sums up the first major challenge for the indexer, which is to: "*try to anticipate what terms people with information gaps of various descriptions might search for in those cases where the record in hand would, in fact, go part way in satisfying the user's information need"(p. 1187)*.

In the model by David, Giroux, Bertrand-Gastaldy, Lanteigne, & Bertrand (1995), thesaurus knowledge is listed as a requirement for indexing, but there is no guarantee the user would use the same keyword as the author/indexer of the document. Even if the thesaurus is contructed with user warrant, different users might use different terms for the same concepts or the useage and meaning of terms might change over time. As indexing typically is a one-time effort, this might results in outdated descriptions.

The discrepancy between indexer and user terminologies is commonly referred to as the semantic gap[10], a disconnect between the intellectual code of the searcher and the indexer/system. While a true bridging of the sematic gap would require a) everyone (users and systems alike) to use the exact same codes for everything and b) a one-to-one relationship to exist between concepts and terms, one solution is using an intermediary. In traditional library settings, the gap could be mitigated by reference librarians, with knowledge of both user-requirements and controlled vocabularies, acting as an interpreter between two different codes. While the reference librarian is not completely gone in a digital library environment, with virtual reference services as a possible replacement, it is not always a realistic or feasible solution, making the problem with bridging the gap as relevant as ever.

Aside from the semantic gap between users and indexers, another gap exists amongst indexers themselves, the so-called consistency problems.

2.3.5 Consistency problems

In 1968, after having spent a decade serving as an advisory editor for the 'Journal of Cataloging and Classification', the American library pioneer Jesse H. Shera held a keynote speech at a conference for the Colorado Library Association. He proposed two laws of cataloging.

Shera's laws:
Law number 1 - No cataloger will accept the work of any other cataloger
Law number 2 - No cataloger will accept his/her own work six months after the cataloging

These laws sum up two very basic problems with allocation of any kind of descriptors to any kind objects - lack of consistency. Referring back to the model of the indexing process and its many kinds of special knowledge, one might assume that following these steps, at least two experts with similar backgrounds would be able to look at the same object (using the same controlled vocabulary) and assign a similar set of descriptors. That this isn't the case is known as inter-indexer inconsistency (fulfilling law number 1) and is a well-documented phenomenon (Markey, 1984). Another issue, also well documented (Leonard, 1977), is intra-indexer inconsistency, the fact that the same indexer might use different terms for the same document at different times (fulfilling law number 2).

[10] The 'semantic gap' was originally used to describe the schism between low-level features of an image (color, texture) and high-level features that are the words describing the image (Smeulders, Worring, Santini, Gupta, & Jain, 2000), the phrase has adopted the broader meaning "*the mismatch between the terms found in documents and those in queries"* (Koopman, Bruza, Sitbon, & Lawley, 2011, p1).

Shera's laws add a certain degree of relativism (at best) or randomness (at worst) to the very principle of indexing/cataloging, which is problematic when the costs related to manual professional indexing is considered.

2.3.6 Traditional indexing conclusion

We have seen traditional indexing as the endeavor by trained professionals to populate fields in catalog records, typically by using ontologies to ensure standardized descriptions, and to lend semantic relations to the descriptors/descriptors they choose. A number of well-known problems, such as the semantic gap, outdated descriptors, lack of consistency and the high price of manual indexing have been mentioned – leading one to consider that the recently emerged alternative to traditional indexing, folksonomies, might be worth considering.

2.4 Folksonomies

Folksonomy is yet another neologism which surfaced during the rise of the social web in the middle of the last decade. The word itself is a contraction of 'folk' and 'taxonomy' and represents a fundamentally different way of thinking about indexing where the wisdom comes from the crowd, rather than the expert.

Folksonomies are part of a new generation of tools for "*retrieval, deployment, representation and production of information, commonly termed 'Web 2.0'*" (Peters, 2009, s. 1) and are the result of users tagging content on collaborative information services.

2.4.1 Web 2.0

The term Web 2.0 was coined by Tim O'Reilly in his much-cited[11] article 'What Is Web 2.0?' in which he presents ideas developed for and during a conference he hosted on the distinguishing characteristics of companies that survived the 2001 dot-com-crash (O'Reilly, 2005

Web 1.0		**Web 2.0**
DoubleClick	-->	Google AdSense
Ofoto	-->	Flickr
Akamai	-->	BitTorrent
mp3.com	-->	Napster
Britannica Online	-->	Wikipedia
personal websites	-->	blogging
evite	-->	upcoming.org and EVDB
page views	-->	cost per click
screen scraping	-->	web services
publishing	-->	participation
content management systems	-->	wikis
directories ("ontology")	-->	tagging ("folksonomy")
stickiness	-->	syndication

Figure 8 – Companies representing the different versions of the web (O'Reilly, 2005)

[11] http://scholar.google.no/scholar?cites=14161329690736649100&as_sdt=2005&sciodt=2000&hl=en

Maness (2006) describes Web 2.0 as "*…a matrix of dialogues, not a collection of monologues. It is a user-centered Web in ways it has not been thus far*", creating a distinction between the "old" web as a one-way publishing channel and the emergence of a more dialogue-based medium. The most prominent examples of 2.0 services are websites like Flickr, MySpace, Facebook, and Wikipedia, that would not exist without their users and the constant update of content they provide.

In a folksonomy the users themselves index the objects by 'tagging' them. So far, the indexing described has been top-down indexing, grounded in the closed world of controlled vocabularies (ontologies) typically found within institutions dealing with classification. A given ontology can have different kinds of justifications – user-warrant being the one that bares closest resemblance to folksonomies. But even when ontologies are created with a specific user-group in mind, the indexing itself is inevitably centered around pre-existing classifications. Referring back to the model of the indexing process, this is tied into the notion of thesaurus knowledge; we choose a set of empty classifications and then go on to populate them with resources.

Even when an indexer, like at KB, does not work with an explicit ontology or some other classification system, there will still exist some sort of consciousness of the wider collection/institution. This becomes apparent when scanning the descriptors selected for the Sven Türck collection. The concepts form clusters (i.e. the indexer intends for certain images to be collocated) – and even though the vocabulary isn't as formally structured as a thesaurus, it is still there and is presumably taken into account in the indexing process. This is also seen in the consistent choice of plural when describing content i.e. even though only one dog appears in the image, the plural 'dogs' is always chosen as a descriptor.

A folksonomy is a bottom-up approach created by social metadata (Ding, et al., 2009) and is resource-centric in that "*Instead of choosing a classification criterion and filling it with resources, it is now the resources that are allocated the criteria*" (Peters, 2009, p. 3) in effect turning the classic approach on its head. The resulting terms (the folksonomy) are in their nature free and messy and lack the different semantic relations (Chapter 2.3.1) found in their controlled counterparts.

2.4.2 Popular folksonomies

This section lists a number of the most popular folksonomies, with an emphasis on the photo-sharing site Flickr. This is far from a comprehensive list, but represents a selection of folksonomies that have been the subject of relevant LIS-research. Newer services like Instagram and Pinterest have not yet appeared in scholarly publications, but could be interesting to look at in the future.

LibraryThing (http://www.librarything.com/): The 'world's largest book club'. A platform for storing and sharing book catalogs and various types of book metadata, including the user generated kind. It is used by individuals, authors, libraries and publishers alike.

CiteUlike (http://www.citeulike.org/): Social bookmarking and sharing for researchers.

Connotea (http://www.connotea.org/): A social bookmarking and reference management site for researchers and scientists. Aside from storing URLs, users of Connotea also have the possibility of store Digital Object Identifiers (DOI), ensuring long term value.

Bibsonomy (http://www.bibsonomy.org/): The 'blue social bookmark and publication sharing system'. Bibsonomy supports exporting of bibliographic data in standardized formats such as EndNote and BibTeX - allowing for easy integration with existing library systems and reference management tools.

Delicious (http://delicious.com/): Another social bookmarking system, with a wider scope and audience than Connotea, CiteULike and Bibsonomy, but sharing many of the same functionalities.

Last.fm (http://www.last.fm/): A popular music sharing website, which allows for end-user tagging of songs, albums and artists.

Youtube (http://www.youtube.com/): Hardly needing an introduction, Youtube had a tremendous impact on the way we watch videos online. Not a very visible feature, but uploaders have the possibility to tag videos.

Flickr (http://www.flickr.com/): The world leading photo sharing site[12]. In April 2011 Flickr reported that more than 6 billion pictures had been uploaded to their servers. Users of Flickr can upload, tag and share pictures and videos via the site.

In early 2006, the National Library of Australia (NLA) enabled public contribution to their image database 'Picture Australia' by publishing their collection on Flickr[13]. NLA was the first National Library to embrace the idea of web 2.0 in this way and two years later, in January 2008, the ball was picked up by Library of Congress (LOC), when they launched a formal collaboration with Flickr, called 'Flickr: The Commons'.

The purpose of The Commons is twofold, as described in the mission statement: "*The key goals of The Commons on Flickr are to firstly show you hidden treasures in the world's public photography archives, and secondly to show how your input and knowledge can help make these collections even richer*"[14].

The Commons is based on the web 2.0 values of user-participation and interaction. Everyone is invited to contribute with input, enhancing the collection via creation of more detailed metadata. When looking at the pictures, the users can either comment on them in the comment-field, favorite them, add a note (identify specific things in the image) or tag them, thus creating an alternative to the controlled vocabulary that we know from the library-world.

Since the inception of The Commons, a multitude of cultural heritage institutions have joined the initiative, hoping to enhance usage of their collections and possibly harvest metadata directly from the users. KB joined The Commons in 2011 and published a small fraction of Sven Türck's photos – which were considered for this thesis, but were omitted due to low number of tags added and the dominance of English-language tags.

[12] According to the web information company Alexa it was amongst the 50 most visited websites in the world in May 2012.
[13] http://www.pictureaustralia.org/documents/MetadataCollectionPolicyweb2_000.pdf
[14] http://www.flickr.com/commons?GXHC_gx_session_id_=6afecb2055a3c52c

Folksonomies come in many different forms and flavors. They can be used to tag any type of content (texts, videos, images and sounds) and it stands to reason that they are as different as the content their users annotate.

2.4.3 Types of folksonomies

Folksonomies are typically differentiated between two different types, broad and narrow (Vander Wal, 2005).

In a broad folksonomy, each object is tagged with the same tag many times by different users. (Lu, Park, & Hu, 2010, s. 769) did a comparison of tags in LibraryThing and Library of Congress Subject Headings and found that the number of tags allocated to a single book can be in the thousands, which in terms of sheer number of terms could actually dwarf the document itself.

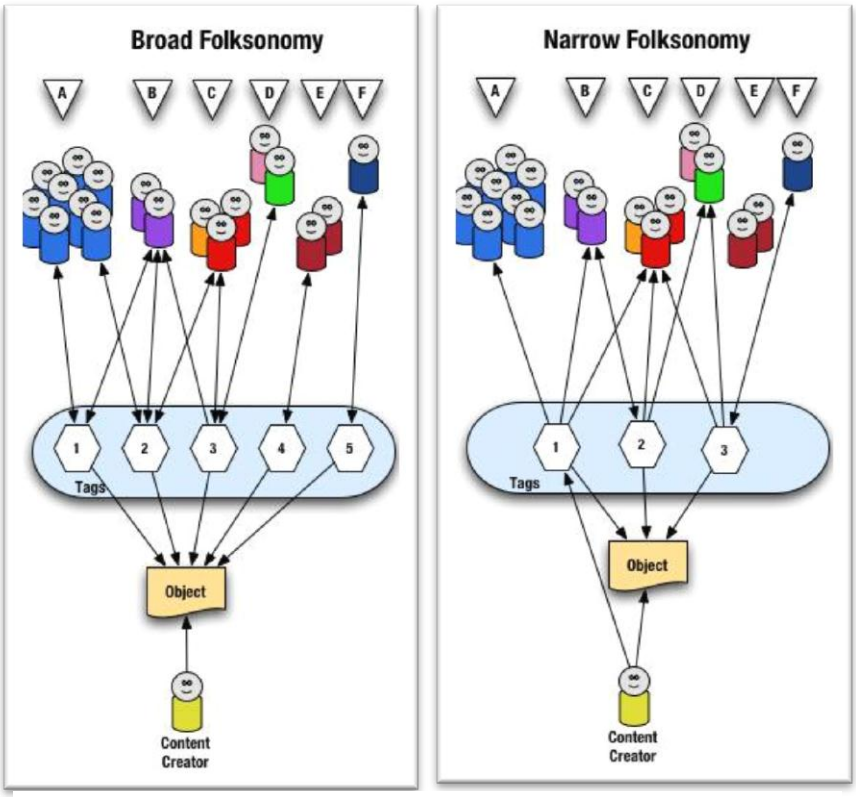

Figure 9 – Broad and narrow folksonomies, from Vander Wal (2005).

In narrow folksonomies tags can only by added once. An example of this kind of folksonomy would be Flickr, where tags cannot be repeated. In this aspect, narrow folksonomies resemble traditional indexing more than their broader cousins.

2.4.4 Tag distribution

Tag frequency in broad folksonomies can be counted on docsonomy (Peters, Schumann, Terliesner, & Stock, 2011) level. Docsonomy is the sum of allocated terms to a single object. Measurements like inverse term frequency can possibly be used to determine relevance – meaning that a tag occuring frequently within a certain docsonomy, but rarely within the entire collection of tags (the folksonomy), is likely to provide stronger discriminatory value along with a high possibility of relevance.

Narrow folksonomies cannot display frequency distributions on resource level, meaning all tags have the same weight/value. It is still possible to determine distributions on folksonomy level i.e. seeing whether certain tags repeat often within the folksonomy. According to (Stock, 2007) the literature provides: "*a large amount of examples for a power-law distribution of tags (and other information units) with its typical 'long tail'*" (p.99).

The power law refers to the fact that a few terms/word will occur much more frequently than all others, a fact not only proven true for tags (Peters & Stock, 2010), but also for other areas of text

statistics, e.g. natural language. In a sample of 28 million words of Danish language, more than 10 million of the words are the same 150 terms[15].

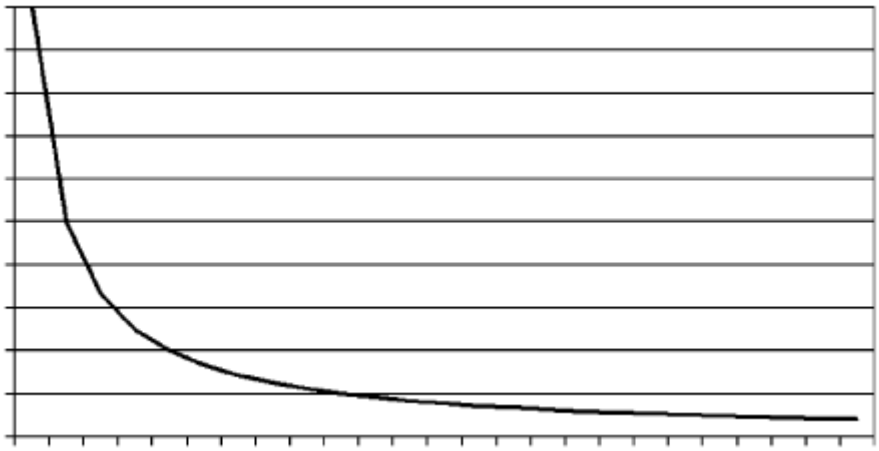

Figure 10 – Typical power law distribution.

The long tail refers to the long slope on the right side of the graph, where all the low frequency tags reside. While these might not constitute consensus in the same way as the more common tags, they might add richness to the description, in that they add unique perspectives and understandings to them. The idea of the long tail was thought up by Anderson (2004) and referred to successful businesses selling smaller quantities of hard-to-find/niche products, rather than large quantities of the same product. On the right side of the graph we find the 'long trunk' where the frequently used terms are.

Another 'power law' documented in the realm of folksonomies relates to the existence of 'power taggers'. The concept both describes the fact that a small minority typically contributes the majority of tags – a similar division of effort has been documented amongst Wikipedians – and the notion that some taggers eventually, after having tagged extensively over time, acquiring knowledge of both the system/platform and the typical vocabulary used, in turn making their tags better. In a report from LOC on Flickr: The Commons (Springer, 2008) it was shown that 40% of the almost 60.000 tags were added by ten of these power taggers (in the report, defined as taggers who provided more than 3.000 tags each).

2.4.5 Tagging process

Unlike the model in figure x of the traditional indexing process, tagging is faster, more intuitive and requires no specialist knowledge. Sinha (2005) provides an overview of the cognitive process behind tagging:

[15] http://korpus.dsl.dk/e-resurser/frekvens150.php?lang=dk

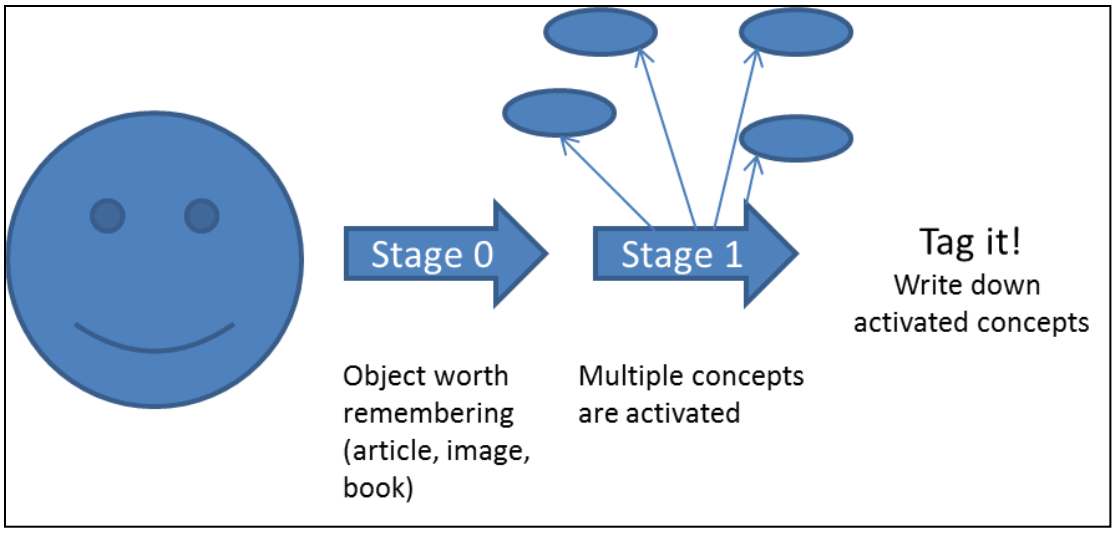

Figure 11 – The tagging process (Sinha, 2005).

The key difference is the lack of choice between concepts: in tagging anything goes.
There is a deep conceptual difference embedded in the purpose/rationale behind the tagging process. Librarians/indexers assign descriptors to facilitate access for other people, i.e. the users of the catalog. A tagger on the other hand, might have several motives/reasons for annotating, of which proving access for other users is only one. Marlow, Naaman, Boyd, & Davis (2006) identifies 'future retrieval' as the most important motivation for tagging, i.e. personal information management, in which tagging is an self-organizing practice similar to structured filing on the users own computer. The other reasons listed are

- Contribution and sharing
- Attract attention
- Play and competition
- Self-presentation
- Opinion expression

Gupta, Li, Yin, & Han (2010) supplement these with:

- Social signaling
- Money
- Technological ease

While Moulaison (2008) creates a top-level distinction between the personal 'exo-tagging' and community oriented 'endo-tagging'.
Unlike the indexing process, the cognitive process behind tagging doesn't require any expert knowledge. Nor does it require the tagger to take the context into account. One simply observes an object and notes down the terms/concepts that seem appropriate – for the multitude of different reasons mentioned above.

2.4.6 Tag categories

Just like we can categorize the motivations behind tagging, we can look at the tags themselves and attempt to categorize them. Tags obviously differ from descriptors, and both the 'linguistic quirks' of tags and the way they differ from standard languages are being researched intensively, along with attempts to identify regularities regarding form and genre within the tags themselves (Peters, 2009, p. 196).

Mathes (2004) studied Flickr and delicious and found that tags could be summerized in eight categories:

- Technical
- Genre
- Self-organization
- Place names
- Years
- Colors
- Photographic Terms
- Ego

Golder and Huberman (2006) identify seven possible functions of tags in delicious:

- Identifying what (or who) a document is about.
- Identifying what the document itself is.
- Identifying who owns the document.
- Refining documents or other tags.
- Identifying qualities or characteristics.
- Self-reference.
- Task organising.

Another attempt to categorize image tags in particular was done by (Beaudoin, 2007, s. 26), by gathering a number of 'power tags' and attempting to create inductive categories by an iterative process of categorization and following validation by user agreement, i.e. test to see whether other people agreed. The highest agreement was in place-names and persons and the lowest in subjective tags categorized as 'rating' or 'poetic'.

Bischoff et al. (2008) did an explorative study of this in more depth across three kinds of objects – bookmarks from del.icio.us, photos from Flickr and music from Last.fm. In the analysis across folksonomies, it was noted that 'Topic' (what or who it is about) was the predominant kind of tag found in both Flickr and del.icio.us, but that a high percentage of 'Location' were unique to photos - probably due to the fact that users of Flickr often tag their own pictures with this type of tag.
These findings were held up against a user study amongst 30 PhD students of computer science asked to rank usefulness/relevancy of the different tag categories on a Likert scale. They found that 'Topic' tags were ranked highest, followed by the subjective category Opinions/Qualities rather than 'Location', suggesting a slight mismatch between image queries and types of tags in image retrieval.

Location was still ranked high on the usefulness scale however, and should not be omitted. The importance of the Opinions/Qualities is interesting and suggests that subjective keywords can be valuable in indexing.

A reductive version of tag categories is provided by Al-Khalifa & Davis (2007):

1) Personal tags for organizing personal object
2) Subjective tags reflecting the user's opinion on the object
3) Factual tags describing facts regarding the object

This provides a top-level system of classification, into which the previously mentioned categories can all be ordered. In the context of a tagging game, the two first categories are fairly useless, as the game mainly rewards the third type of tag (even without a detailed knowledge of the game, this strategy seems fairly self-evident).

Regarding distribution among word classes Heckner, Mühlbacher, & Wolff (2008) reports 72% nouns, 12 adjectives, 15% acronyms and 1% numbers for connote-tags, while a different study found 90% of all tags in Flickr to be nouns (Guy & Tonkin, 2006).

One conclusion to be drawn, is that the tags (and the categories one can derive from them) are as heteorogenous as the different folksonomies they come from, but some common traits can be noted and can thus be used to inspire the categorization of tags and descriptors in the following chapters.

Tags can be either personal (user-dependent) or be extrinsic to the tagger and his/her relationship to the object (user-independent). The latter types also include subjective tags, as they can have potential meaning for other users when searching or browsing. In the context of a game, tags for personal information management are presumed to be non-existant, as the users have no possibility to use the tags outside the game and no motive for adding those types of tags.

2.4.7 Tag navigation

A folksonomy is essentially a flat space of keywords, lacking the semantic connections known from the ontologies (Hassan-Montero & Herrero-Solana, 2006). There are, however, ways to make them better for navigation, like the generation of a tag cloud or other visual representations. The user clicks the tag and is lead to all objects within the system that have that particular tag attached to it.

This type of exploratory browsing is called 'pivot browsing' and it allows users to make serendipitous finds by stumbling upon relevant/surprising terms or phrases (Peters, 2009).

I have used the free tag cloud generator TagCrowd (http://tagcrowd.com/) to generate a sample cloud by using the 30 most frequently occurring terms across all four sets of data:

Figure 12 – Tag cloud for Sven Türcks collection.

The software had some problems displaying the 'æ,ø,å' characters from Danish, but provides an example most users of collaborative information services will instantly recognize.

2.4.8 Folksonomies conclusion

Jäschke, Hotho, Schmitz, Ganter, & Stumme (2008) write that, unlike ontologies: "folksonomies do not suffer from the knowledge acquisition bottleneck, as the significant provision of content by many people shows" (p.38). The 'bottleneck' is a concept derived from knowledge management literature. It encapsulates the problems related to sharing of knowledge/information and the main components (Wagner, 2006) correspond well to the challenges/problems attributed to indexing in the previous section:

- Narrow bandwidth (scarcity of resources). *Indexing by professionals requires expensive training and time when performing the task.*
- Knowledge inaccuracy (consistency problems). *Experts/indexers make mistakes and cannot always be expected to perform better than average users.*
- Acquisition latency (scarcity of time) *Holley (2010) mentions that photo archives suffer from a severe lag, as digitized images await metadata.*
- Maintenance trap (outdated vocabularies) *Creating controlled vocabularies requires both resources and continuous updates to reflect the actual language used by people using the systems.*

Folksonomies addresses these issues, being both cheap, fast and possibly closer – they represent a potential alternative to traditional indexing, but come with their own weaknesses as well.

Folksonomies like we know them from Flickr or Delicious weren't intended to be folksonomies, but primarily ways to organize personal information. I already described how the motivation for tagging mainly is for personal purposes, and this is a key difference between the two kinds of knowledge organization systems. LibraryThing is probably the purest combination of the two, in that it is a community based effort specifically aimed towards bibliographic description to be shared with others. Flickr: The Commons is another example of this convergence of intentions.

The overall pros and cons are summed up by Mathes (2004): "*Its uncontrolled nature is fundamentally chaotic, suffers from problems of imprecision and ambiguity that well developed*

controlled vocabularies and name authorities effectively ameliorate. Conversely, systems employing free-form tagging that are encouraging users to organize information in their own ways are supremely responsive to user needs and vocabularies, and involve the users of information actively in the organizational system".

One more thing is the simple fact that ontologies/controlled vocabularies aren't practically or economically extensive, meaning that folksonomies are the only way to index the web, as human indexing won't be able to cope with the vast amounts of information online (Quintarelli, 2005). This last argument was purveyed spot-on by the Dutch artist Erik Kessels in his 2011 art-installation that consisted of a room filled to the brim with printouts of every single picture uploaded to Flickr within a 24-hour period:

Figure 13 – Erik Kessels visualization of Flickr

That being said, there might be valuable lessons to learn for both types of indexing, by comparing them with each other.

2.5 Overlap between tags and descriptors

As mentioned in chapter 1.5 a substantial amount of research has been done on the relationship between ontologies and folksonomies. Of special interest for this thesis are the various investigations into the overlap, syntactic as well as semantic, between tags and descriptors.

- **Syntactic overlap** is defined as matching terms, either by exact match (character-by-character) or by fuzzy match, which takes orthographical variations into account and thereby matches word-stems.
- **Semantic overlap** describes whether the 'meaning' behind the two types of keywords can be said to be linked for instance by thesaurus relations.

Al-Khalifa & Davis (2006) measured the syntactic overlap between folksonomies and indexers and found there to be a 19.48% overlap between the tags and descriptors.

Kipp (2005) looked at semantic overlap (see 3.3.3 for more on this) between author keywords, descriptors and tags and found that the most frequent semantic overlap among 1342 terms assigned to 165 articles were associative relations (340), followed by syntactic overlap of 155. Almost half (573) were not related. She does distinguish clearly between overlap between her three datasets, which is why percentages aren't included.

Wetterström (2008) did an experiment in which a small group of taggers were asked to tag 217 documents to determine the semantic overlap between the existing descriptors (LCSH) and the tags. He found that the exact syntactic overlap between tags and descriptors was 9.14%, but that broader and narrower terms constituted 14.61% and 19.62%, respectively, of all the tags.

Lykke, Hoj, Madsen, Golub, & Tudhope (2012) used semantic overlap as a metric to determine whether a tag recommender system would change the semantic overlap between tags and INSPEC descriptors. While they didn't find large deviations between tags coming from a recommender-free and a recommender-enabled system, their findings are still of interest to this study; they only found 3,0% (recommender-free) and 4,4% (recommender-enabled) syntactic overlap between tags and descriptors. An interesting point here is also the 'related term' overlap, defined as being some sort of associative relation. Every single tag, not fitting into any of the other thesaurus-categories was placed in this category. This is interesting when considering what actually is meant by 'associative relation' – a discussion I will continue in my methodology (chapter 3.3.3).

Thomas, Caudle, & Schmitz (2009) aggregated a number of different folksonomies (Librarything, PennTags, Aquabrowser, SOPAC, MTagger & Encore) and compared the tags found there with the assigned LCSH for the same books. They found an exact overlap of 6% and very low values for broader/narrower terms (0.3% for both) and similar low values for related terms. Addition they noted that 22% of the tags were somehow (but not formally) related to the existing descriptors.

Overall, the results from these different studies diverge quite substantially, but a common finding is the rather low syntactic overlap, which ranges from a maximum of 19.48% to a minimum of 3.3%. Keeping the problems regarding indexing (inconsistencies) in mind, these results are hardly surprising.

To the best of my knowledge, there has been no study of semantic overlap within the realm of image indexing so far, which is interesting, as image indexing comes with its own unique challenges.

2.6 Images

Enser, Sandom and Lewis (2005, s. 178) provide a taxonomy of still images to give a detailed overview of the types of images which can be found in image collections.

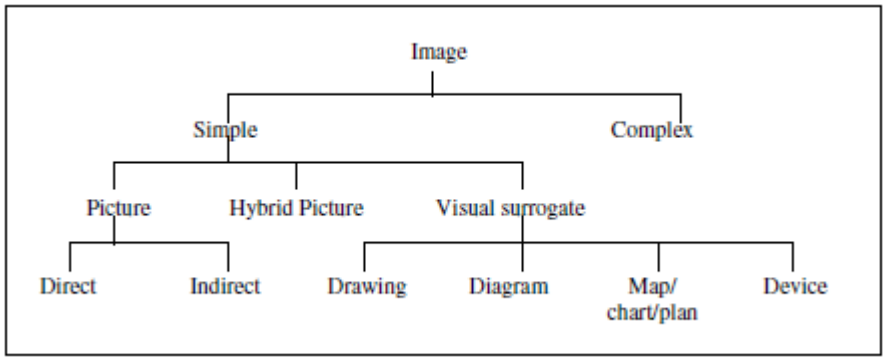

Figure 14 – Taxonomy of still images

Each types of image can be investigated with indexing in mind. Collections of old maps make a good example, as they will often be heavily annotated, presenting us with an interesting challenge: are the best indexing terms text lifted directly from the map?

In the case of Sven Türck, we are dealing with direct pictures. While the direct pictures are the most common ones dealt with in the literature, is should be noted that significant parts of the collections at KB are old maps and that almost half of the current publications in the new database are hand-drawn illustrations from old newspapers.

2.6.1 Image interpretation

An oft cited theory when it comes to interpretation of images is given by the French art historian Panofsky (1970), which affords his division of viewpoints, representing different attributes in an image and different levels of interpretation:

- Pre-iconographic or primary subject matter.
- Iconographic or secondary subject matter
- Iconological or intrinsic meaning

In the case of a picture of the little mermaid, we are looking at:

- Pre-iconographic: Statue, Ocean
- Iconographic: Little Mermaid, Langelinie
- Iconological: Melancholy, Loneliness

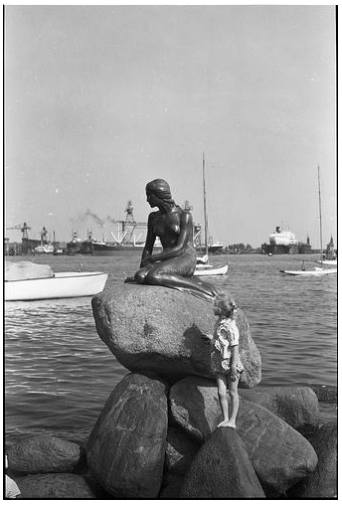

Figure 15 – Little Mermaid portrayed by Sven Türck

Where the Iconographic level qualifies the very general terms from the Pre-Iconographic level, in this case turning them into proper nouns for a much higher degree of specificity (and probably also creating better keywords by doing so), the Iconological level is based on interpretation and is therefore subjective. In this case I added the keyword “Melancholy” because I always found the

statue and the story it represents very sad. On the two first levels of interpretation, we can form a consensus on whether the keywords are relevant/true or not. Other people, that might only know the 'Disneyfied' version of Hans Christian Andersen's original fairytale, might have completely different associations with the little mermaid and her story.

The most important/problematic distinction is between Iconographic and Iconological as it is here we enter the realm of interpretation. Panofsky's division echoes throughout the literature and is also referred to as the ofness (concrete and objective) and aboutness (abstract and subjective) (Layne, 1994), hard (what is *in* the image) and soft (mere ethereal aspects of the image) (Krause, 1988) or denotation (what is in front of the lens) and connotation (the meaning and function of the image) (Yoon & O'Connor, 2010).

2.6.2 Image indexing

When it comes to querying for images, there are two approaches, graphic queries and textual queries (Ménard & Arsenault, 2008). Graphic queries are visual input a system matches to existing images in the database by some criteria (color, shape, texture) and returns relevant hits. Such systems are called content-based image retrieval systems, or CBIR, systems. A recent example is tineye.com[16], a search-engine that indexes images found on the web. A graphic query (in the form of a file, such as a JPEG or a URL directing to a picture) in tineye.com returns all other images from the index with a similar pixel-structure. This means that a user can find higher resolution versions of images or check if a picture really is as original as someone claims it to be. Another system is ALIPR[17], where users can upload images and help the system understand patterns. But these CBIR systems are not widespread and are still in their infancy, and the most common way to search is through textual queries.

The two methods of querying are reflected in the two types of indexing, Content Based Image Indexing and Concept Based Image Indexing. The first, also displayed in figure 5 (chapter 2.2), relates to the picture 'as is it'. By means of automatic analysis, computers attempt to decode an image (Smeulders, Worring, Santini, Gupta, & Jain, 2000). This might be easy for colors or simple patterns, but moving beyond pre-iconographic descriptions presents a computer with significant challenges, such as describing a mood, identifying a location or interpret meaning (Enser P. , 2007), which is why the reliance on 'concepts' still holds relevance.

Ménard (2007) states that: "*When using textual queries the success of the retrieval largely depends on the correspondence between the query of the searcher and the text associated with images"* (p. 91) which cuts into the heart of the matter. The challenges when "*trying to translate visually coded knowledge into a verbal surrogate*" (Enser, Sandom, Hare, & Lewis, 2007, p. 466) can be enormous. Especially when we have to make ends meet, i.e. making sure that we translate in the same way. A familiar way to describe this challenge is when playing the popular family-game 'Pictionary'[18], which recently had a tremendous renaissance in the form of the best-selling app for smartphones and tablets 'Draw something'. The core of the game is that a player draws a card with a term or a phrase, which he or she in turn has to illustrate. The team-members then try to guess the correct term.

[16] http://www.tineye.com/about
[17] http://alipr.com/
[18] http://en.wikipedia.org/wiki/Pictionary

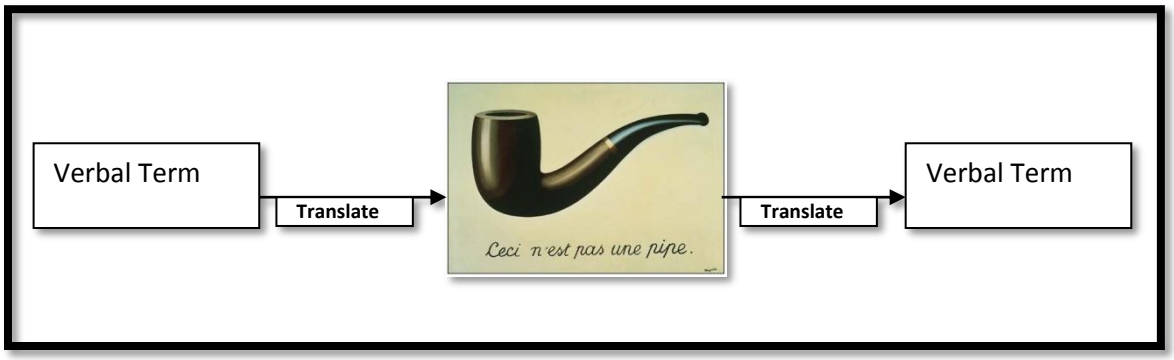

Figure 16 – Pictionary as a process.

Sometimes it is easy. Most of us can draw a dog or a ship, or a pipe for that matter. But when the visually coded knowledge is more complex than one simple object, we need more words to describe them and the translation in itself might become increasingly ambiguous. One term will rarely (if ever) provide an exhaustive description of an image.

When playing Pictionary you have the added advantage of body language and oftentimes some common frame of reference. When searching a database for an image using textual queries, the computer only returns exactly what you ask for. No hints are given and nonverbal communication is nonexistent. A picture is worth a thousand words, but without words, you'll never find it.

Even if you find the perfect term(s) and produce the best possible query, it is not given that whoever 'translated' the visual data in the first place, did it the same way you would have – again noting the semantic gap. Shatford (1986)observed how "*the delight and frustration of pictorial resources is that a picture can mean different things to different people*" (p. 42).

Viewed in this light, it comes as no surprise that indexing consistency studies in image collections yield as poor results as those done for other types of materials. Hughes & Rafferty (2011) did a study on inter-indexer consistency between two indexers from the National Library of Wales – both using the Thesaurus for Graphic Materials (TGM) – who only achieved approximately 27% consistency. Markey (1984) used three indexers to index the same images and found consistencies ranging from 1-27% depending in the images in question.

2.6.3 Image term categories

As RQ1 revolves around term-categories, this section is dedicated to exploring the different relevant frameworks for image term categories.

Jörgensen (1998) developed a framework for describing image attributes by letting 48 masters students perform describing tasks of six images. She reported the following term categories (p.169):

- **Object** Objects that are visually perceived, e.g., body parts, clothing
- **People** The presence of a human form
- **People-related attributes** The nature of the relationship among people, social status, or emotions

- **Art historical information** Information related to the production context of the image, e.g., artists, medium, style
- **Color** Specific named colors or terms relating to various aspects of color
- **Visual elements** Elements such as composition, focal point, motion, shape, texture
- **Location** Both general and specific locations within the image
- **Description** Descriptive adjectives (e.g., wooden, elderly), size, or quantity
- **Abstract concepts** Attributes such as atmosphere, theme, or symbolic aspects
- **Content/story** A specific instance being depicted
- **External relationships** Relationships to attributes within or without the image, e.g., similarity
- **Viewer response** Personal reaction to the image

She stated that social status and activity are more interpretive and can be said to be related to the story of the image, i.e. what is happening in it. The most used term-categories were object, people, color, content/story and location. Jörgensen (1998) herself pointed out that the results are only *"…suggestive rather than conclusive. While many procedures were adopted to increase reliability and validity of the data, the research remains exploratory and these distribution cannot at this point be interpreted as statistically significant*" (p.168). But she does say with some certainty that 'primary visual content' i.e. the things *in* the picture are more likely to be described.

Jörgensen's framework for image descriptions has been widely applied by image researchers when analyzing indexing of images (Rorissa, 2010). Enser and McGregor (1992) analyzed a large number of image queries and classified them in four categories:

- Unique
- Unique with refiners
- Non-Unique
- Non-Unique with refiners

Where the 'uniqueness' refers to named people, objects or places (iconographic) and refiners typically were a way to specify the term e.g. with a year. Non-uniqueness entails more generic terms (pre-iconographic).

In an analysis of image queries in the field of art history, (Chen, 2001) combines the two and conveys that, the only categories from Jörgensen she found repeated in her data were Location, People and Literal Objects. Furthermore, a distinction between unique and non-unique locations was made; queries were made for unique locations, i.e. proper nouns.

The two studies mentioned last (Enser & McGregor, 1992; Chen, 2001) are incidentally examples of the request/query-based indexing – by discovering what types of terms the users are actually using when performing image retrieval, indexing practice might be improved to match their information need better.

In her PhD Dissertation on press photos Ørnager (1999) notes seven objective and one subjective term categories[19] among keywords from a photo archive:

- Places (objective)
- Scenarios (objective)
- Objects (objective)
- Topic (objective)
- Action (objective)
- People (objective)
- Year (objective)
- Mood (subjective)

When using the categories to analyze attributes in St. Andrews Library Photographic Archive, (Rorissa, 2010, s. 2235) found the following term category distribution for descriptors:

- Location (43.35%)
- Content/story/event (25.48%),
- Object (22.51%)
- People (4.83%)

And a different distribution for Flickr Tags:

- Location (30.63%)
- Content/story/event (17.14%),
- Object (10.51%)
- People (14.98%)

Location also encompasses place-names, which occur frequently in the data. Her findings show how a quarter of the index terms actually relate to what is happing in the images – which could mean that a lot of interpretation is occurring in the indexing phase. As described later (chapter 3.3.1), these studies informed the term categories in this thesis.

An alternative to the image term-categories could be a facet-analysis and comparison of which facets tags and descriptors cover. This was attempted by Conradi (2011), but the researcher found the methodology to be a *"highly laborious and time-consuming effort"* (p.21) and this dimension was therefore left unexplored.

2.6.4 Image conclusion

As evident from the complexities covered in this section, indexing of images by some sort of verbal description is likely to be even more subjective than it is when indexing texts (Lancaster, 2003, s. 217).

[19] Translated from Danish

This knowledge led (Brown, Hidderley, Griffin, & Rollason, 1996) to suggest a more 'democratic' approach to image indexing in which users, not indexers, provide the keywords; a precursor to the now-widespread phenomenon folksonomies.

Another, more active, way to collect metadata/descriptions from the users themselves, is to aim specifically for that type of output by constructing either a service that facilitates or promotes it (e.g. by allowing tagging in an OPAC) or designing a Human Computation Game, as happened at KB, which can be considered crowdsourcing.

2.7 Crowdsourcing

Since his 2006 article (mentioned in 1.1.), Jeff Howe has maintained a blog on crowdsourcing, in which he defines crowdsourcing as:

"the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."

The eponymous 'crowd' is supposedly able to perform the job better than the single employee – 'better' meaning cheaper, faster and possibly more reliably, depending on the kind of project in question.

The core idea behind the wisdom of many was first put forth by Galton (1907). He conducted an experiment in which a group a people was asked to guess the weight of a bull. By adding all the guesses and calculating the average, he found that the mean value actually was better than any of the individual guesses. This serves as an apt metaphor for the strenght of folksonomies; in the context of knowledge organization/representation, one can assume that folksonomie are semantically 'richer' than indexing done via ontologies, since more opinions and perspectives are taken into account.

During the Distributed Community Empowerment 2011 EuropeanaTech conference, various exponents of this notion were presented and an idea, which was repeated several times, was the concept of the 'cognitive surplus' from the 2010 book with the same name by Clay Shirky. The basic questin posed by Shirky is: what if we, as a species, focus all the time/energy we spent on watching TV on endeavors more 'worthwhile', such as Wikipedia?

He underlines the notion with a very poignant illustration of our cognitive surplus as seen in figure x. If all US adults direct all their TV-watching time toward collaborative efforts, that are made possible by modern IT, we would be able to produce 2000 Wikipedia-like projects annually. Another spin on this vast

Figure 17 – Cognitive surplus (Shirky, 2010)

potential resource was given by von Ahn & Dabbish in their first article on Human Computation Games (2004): "*If the game is played as much as popular online games, we estimate that most images on the Web can be labeled in a few months*."(p.319)

This statement predates Flickr and Facebook and the ocean of images now available online (as seen in 2.4.7), but it still represents an enticing idea.

While the term crowdsourcing is relatively new, the concept itself has quite a history. The New English Dictionary, which was created several centuries ago, was to a large extent made possible by civilian donations – volunteers that found words in the wild and sent them and definitions to a board of editors. Another example, that significantly predates any notion of IT, is the annual Christmas Bird Count organized by the National Audubon Society that in 2011 saw more than 62.000 thousand volunteers tally more than 60 million birds across North America, Canada, Latin America, the Caribbean and the Pacific Islands[20] , and began in 1900 with just 27 volunteers.

The latter example is actually a subgenre within crowdsourcing called Citizen Science, in which researchers use volunteers to perform otherwise unobtainable measurements. Aside from Citizen Science, a lot of other subgenres occupy the sprawling crowdsourcing landscape, as rendered by Dawson (2010):

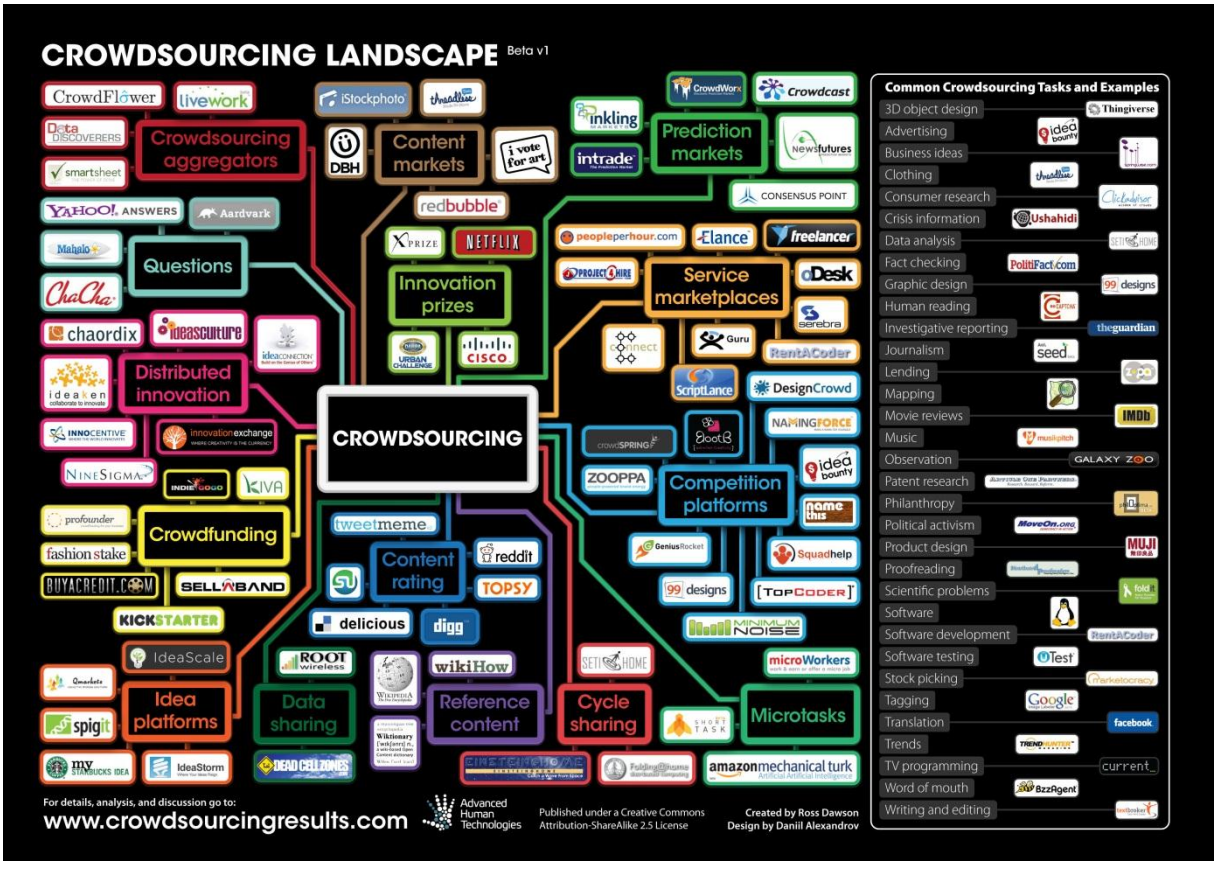

Figure 18 – Crowdsourcing landscape by Dawson (2010)

[20] http://www.audubon.org/newsroom/press-releases/2011/audubon-s-112th-christmas-bird-count-reigns-model-crowd-science

2.7.1 Notable examples

This section lists a few interesting examples of crowdsourcing.

reCAPTCHA

Figure 19 –Example of the reCAPTHA mechanism

Known to most people with an internet connection, reCAPTCHA provides protection against bots trying to access restricted parts of websites, but not everyone may know that the distorted text the users are asked to type in, is actually a crowdsourced OCR-reading of digitized books and newspapers. The program was originally developed by the same Luis von Ahn behind the first Human Computation Games and is now owned by Google[21].

Galaxy Zoo[22]

Belonging to the citizen science subcategory, this project revolves around classification of galaxies, a task computers cannot perform. Within their first year, more than 50 million galaxies were classified by almost 150.000 people.

Mechanical Turk

An 'artificial artificial' intelligence run by Amazon, it serves as a marketplace for micro-transactions; the users (the Turks) perform small tasks, for which they receive money.

The Descriptive Camera[23]

In the spring of 2012, an interesting utilization of the Turks was shown by a New York University student Matt Richardson in a class on Computational Cameras; he created a camera that instantly uploaded the image to the Mechanical Turk and short thereafter printed out a natural language description by a Turk, somewhere in the world. Unlike a folksonomy like Flickr where multiple taggers can provide their viewpoint, the camera is the result of a single observation, thus limiting the output to a single anonymous individual; it does still have merit in terms of pure speed and possibly cost efficiency (an image description created by the camera costs $1.25).

Copenhagen City Archives[24]

In an ongoing transcribing project at Københavns Stadsarkiv (Copenhagen City Archives), they have recently passed one million addresses transcribed. It is the result of approximately 400 volunteers contributing a total of twenty Full Time Equivalents (roughly 40.000 hours) so far. Københavns Stadsarkiv itself has used only two Full Time Equivalents from their budget on starting and maintaining the project, making a rather strong argument for user participation in a cultural heritage context.

[21] http://www.google.com/recaptcha/faq
[22] http://www.galaxyzoo.org/
[23] http://mattrichardson.com/Descriptive-Camera/
[24] http://www.politietsregisterblade.dk/

Authormagic

Within the field of High Energy Particle Physics, the CERN-developed system Authormagic allows authors to identify (disambiguate) themselves and thereby create their own authority files. This can be very useful in a discipline were co-authors can number in the thousands (Brooks, Carli, Dallmeier-Tiessen, Mele, & Weiler, 2011). A similar effort is described by Bainbridge, Twidale, & Nichols (2011) in developing a system for crowdsourcing name authority files, mentioning the 'usual' problems of traditional catalogs i.e. costly, not-updated and despite best intentions, erroneous.

Waisda? http://www.cs.vu.nl/intertain/2011/10/waisda-video-tagging-game/

A Human Computation Game, also mentioned in the introduction. Mimics the basic gameplay of the original ESP-game, just like GeF, but uses it for annotation of video, rather than images.

Digitalkoot http://www.digitalkoot.fi/en/splash

A Finnish Human computation game in which players take the role of a mole and help it pass bridges by correcting OCR-scanned words from the libraries digital collections. If the word is verified by other players, the bridge becomes more solid.

2.7.2 Crowdsourcing in libraries

Crowdsourcing in a cultural heritage context can serve a multiple purposes. Aside from the rationalization/cost and the different way the crowd can accomplish things single indexers/institutions cannot - there is another benefit in engaging patrons in some sort of activity, either in 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.

Modern libraries have to various degrees relied on volunteers and the surrounding community for various favors. Crowdsourcing is not a new idea, but one that has dramatically changed with modern technology; Wikipedia could have been a done by writing letters, but the logistical challenges would be insurmountable, and modern information technology has enabled large-scale, as well as smaller-scale, projects to emerge.

Holley (2010) lists the ways in which libraries could benefit from crowdsourcing:

- Getting users to mark the errors in our catalogues
- Rating the reliability of information/records
- Adding information to records
- Verifying name authority files
- Adding user-created content to collections
- Creating e-books
- Correcting full text
- Transcribing handwritten records
- Describing items that have not been made accessible because they are not catalogued/described yet.

She puts special emphasis on the last point, and mentions how: "*A prime example of this is photographs. The normal procedure in a library is that a photograph is not digitised until it has been catalogued. If instead it is digitised first and users are given the chance to describe the content this would radically open up access to a lot of 'hidden' and difficult to describe photographic collections*" (Holley, 2010).

Oomen & Aroyo (2011) provides an overview of crowdsourcing research and projects from the cultural heritage domain and identifies the following

- Correction and Transcription Tasks
- Contextualisation
- Complemention Collection
- Classification
- Co-curation
- Crowdfunding

Only diverging from Holley (2010) with the addition of 'crowdfunding' – referring to the "*Collective cooperation of people who pool their money and other resources together to support efforts initiated by others*" (Oomen & Aroyo, 2010, p.3)

GeF belongs to the 'Adding information to records' and 'Classification' categories, and while the Sven Türck collection, wasn't made available for tagging before professional indexers assigned descriptors like Holley suggests, the creation of GeF is a step in that direction.

2.7.3 Crowdsourcing motivations

Müller, Thoring, & Oostinga (2010) identifies four distinct reasons for users to participate in crowdsourcing activities: money, altruism, usefulness, and fun.

While monetary rewards doesn't seem like a likely incentive, Holley (2010) points out, that libraries already are proficient in getting people engaged in their activities. The very idea of a (free and public) library and the freedom of and equal access to information seems highly altruistic and the notion of usefulness is also tied to these ideals. Introducing games as tools for crowdsourcing could seemingly be a way for libraries to add 'fun' to the equation.

Owens (2012) states that most successful crowdsourcing projects aren't actually about crowds at all, but rather about smaller scale volunteerism and citizen involvement in creating public goods, for which cultural institutions already have a long standing tradition. He argues that crowds in most cases are small communities of self-motivated individuals; something we also saw was true on Flickr: The Commons, with the small inner circle of 'power taggers'.

This point was also raised by in an interview with GeF co-creator Tom Juul Andersen (2010) in which he points out that GeF is only the beginning and that KB in the long run will attempt to focus their crowdsourcing-projects to more specific audiences by doing targeted marketing towards small communities with a special interest in particular collections e.g. by getting train-enthusiasts to tag pictures of trains.

2.7.4 Human Computation Games

Human Computation Games or Games With A Purpose (GWAP) was pioneered by Louis von Ahn and Laura Dabbish in 2004 (von Ahn & Dabbish, 2004) with the ESP-game, as a way to address the image

labeling challenge. Introducing the idea they write: "*Rather than using computer vision techniques, which don't work well enough, we encourage people to do the work by taking advantage of their desire to be entertained"(p.319)* subscribing to the idea of 'textual queries' and 'concept-based indexing'. Since automatic indexing isn't yet a feasible option, crowdsourcing in the form of a game could be one.

GWAP has since then developed into a sort of running experiment in Human Computation Games, with a wide range of games (of which the ESP-game was the first) and media types involved (Law & von Ahn, 2009). They are typically very simple, fast and intuitive.

Yuen, Chen, & King (2009) establish a taxonomy distinguishing between four fundamental classes of Human Computation Games:

- **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 GeF-game, like the Dutch Waisda? and the original ESP-games are all examples of 'Output-agreement games.

In their original 2004 introductory article to GWAP von Ahn & Dabbish did a manual evaluation after the first 1000 pictures had been thoroughly tagged, by asking 15 participants to evaluate the tags. They found that 85% of the validated tags were deemed as 'useful when describing it' (von Ahn & Dabbish, 2004).

For the reasons mentioned in chapter 1.5 player generated tags have not before been compared to indexer generated descriptors.

2.7.5 Crowdsourcing Conclusion

While the two concepts of folksonomies and crowdsourcing are closely related, they are different in many ways. Crowdsourcing is a conscious effort by someone. The output can be a folksonomy, but it can also be an encyclopedia, OCR reading, name author error correction, galaxy classification or even the cure for diseases. The main point is that crowdsourcing methods are as diverse as the problems they are created to address. If a library wishes to tap into the 'cognitive surplus' they should launch services which create incentives for the users to participate. There might also be a point in catering to a smaller audience of power-users, rather than trying to reach everyone.

2.8 Chapter summary

This chapter has described indexing and the problems with assigning descriptors to objects in general and images in particular. A radically different approach to indexing, which has emerged within the last decade, the folksonomy, was furthermore explained.

An institutional alternative or supplement to traditional indexing might be to harness the wisdom of the crowd. This can be done by utilizing existing collaborative information services - the folksonomies – but can also be an active effort, in which the institution partakes more actively, by setting up systems which allow for more user-generated content.

The Human Computation Game GeF is an example of this, and the output of this game can be considered as a folksonomy – just like the original descriptors can be considered to be exponents of traditional indexing by use of an ontology – and the these two types of metadata will be the units of analysis in this thesis.

Chapter 3 Methodology

3.1 Chapter overview

This chapter is divided into two overall sections, one smaller section on data collection and one on research design.

The data collection is divided into a part about procuring the relevant literature used throughout the thesis, and a part relating specifically to the technical aspects of collecting/modifying the tags and descriptors I used as my data.

The research design is split into a part for each research question, as answering them requires different methods. I will describe the methods employed in my analysis as well as a justification for my choices.

While the method employed to answer RQ2 is quantitative, the first and last research questions are more qualitative in nature. As the analysis is done by text (term) reading and categorization, there is a shift from the technical/objectivist towards the more interpretive/subjectivist end of the continuum.

Probably the most fundamental operation in the analysis of qualitative data is that of discovering significant classes of things, persons and event and the properties that characterize them. By relying on earlier studies in the initial phase (both regarding the term-categories addressed in RQ1 and the thesaurus-relationships addressed in RQ3) allowed for a considerable amount of time to be saved.

Another approach could have been a complete 'tabula rasa' categorization, but standing on the shoulders of the previous research done in the field on both term-categories and thesaurus-relations seemed like the most fruitful approach.

3.2 Data collection

3.2.1 Literature search

Indexing & image indexing, crowdsourcing & folksonomies and Human-Computation Games constitute three different subjects, each requiring a different information searching strategy.

Indexing has been investigated thoroughly in the literature. It is well-known concept and has been approached from a multitude of different perspectives. In such a case, I chose monographs rather than articles as my starting point. By selecting a few choice sources (Lancaster, 2003 & Svenonius, 2000) I was able to get exhaustive coverage of indexing. For image indexing I chose two comprehensive literature reviews as my foundation (Rasmussen, 1997 & Enser, 2007) and branched out from there.

Crowdsourcing and *folksonomies* on the other hand, are much newer concepts. Aside from searching for the two terms in LISA and LISTA, I chose a list of high quality journals in the field of library and information science and scanned abstracts for the years 2005-2011 to identify relevant sources. I

chose the highest tier journals from discipline 27 "Library and information science" in the 2011 list[25] compiled for the Danish Ministry of Science, Innovation and Higher Education by an expert group in the field. In addition, I used a PHD thesis by Peters (2009) to find more sources on folksonomies.

Human Computation-Games is very specific concept, invented by a single author, Luis von Ahn. His original 2004 "Labeling images with a computer game" from the SIGCHI conference on Human factors in computing systems was used as the basis for a 'citation pearl growing' search in Google Scholar. This resulted in a manageable 86 articles to scan for relevant information.

3.2.2 Collection of descriptors and tags

In order to be able to work with the data in a systematic way I had to compile the existing descriptors alongside with the tags and normalize the two different datasets in a spreadsheet to accommodate speedy analysis. I will describe the methods used in the following.

3.2.2.1 Descriptors

The pictures in Sven Türck have already been published online by KB via the proprietary software Cumulus Online Publishing. The metadata is accessible via KBs OAI-server allowing for harvesting in both MODS and Dublin Core (both in XML) by everyone with knowledge of OAI-requests and a browser, but even though the Sven Türck collection is available, it is not a unique set. The photos are a part of a larger set, "Billeder", which contains over 37.000 records, so the entire "Billeder" set was harvested and stored locally. The repository only provides 1.000 records at a time, which resulted in 38 separate XML files of metadata. Though the metadata could have, in theory, been merged into a single file, the hardware and software in use would have been unable to cope with such a massive file. A sample file in the original MODS metadata is in Appendix A. In order to isolate the Türck collection, XSLT was used to extract relevant metadata from only those records in which Türck was the creator. To avoid applying the XSLT to the files 38 separate times, a 'master' XML file was created which referenced all the files at once. Appendix B includes a sample from the master file, the XSLT used to separate the Türck photos, and a portion of the resultant XML file. Finally, XSLT was used to format the Türck records into an HTML table, which was copied into Excel (see Appendix C). It should be noted that descriptors are taken from several different MODS elements, including the general descriptors and different subject fields e.g. person, subject and location. As the raw MODS has a lot of redundancy on object level, i.e. geographic location is mentioned in both general descriptors and locations, de-duplication was done on object level, to clean the data for analysis.

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, the descriptor is omitted for the analysis. 'Denmark' is meaningless as a search term, as it will result in almost total recall of the entire collection, i.e. it does 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. The two-word descriptors in question were the following:

[25] http://www.fi.dk/viden-og-politik/tal-og-analyser/den-bibliometriske-forskningsindikator/autoritetslister-for-tidsskrifter-og-forlag/Autoritetslisten%20for%20tidsskrifter%202011%20-%20med%20niveauer.pdf

slotte og herregårde[26] (129)
ferie og fritid[27] (172)
jul og juleskikke[28] (40)
industri og håndværk[29] (26)
kirker og kirkegårde[30] (70)

A total of 437 compound descriptors were treated this way, adding 6.3% to the total number of descriptors, bringing the final number of descriptors in the analysis up to 7306.

3.2.2.2 Tags

Thanks to connections at KB, I was able to get a 'raw' dump of the data collected via the Facebook-game. This dataset simply consisted of a number of URI's (objectIDs) in one column and a list of free tags (22.787) in CSV-format for excel. Another dump was provided with the 2516 3Vtags, i.e. tags which three or more users had provided. While the validation threshold normally is set at three, I quickly realized that a lower threshold (2Vtags) might provide a richer dataset (see chapter 1.5).

In order to include this in my analysis to get the tags that appeared two times on object level, the Excel table was exported as an XML file to make manipulation with XSLT possible.

The first XSLT included a tag only if it contained the same value as a previous tag for the same record. This eliminated all the tags that only appeared once, but also resulted in duplicate tags for any that appeared 3 or more times. So, another XSLT was used to remove any duplicate tags. This could have been streamlined into a single transformation, but the first XSLT was created before the error was noticed, so it was simpler to create a second XSLT to correct the problem. Appendix D shows both of the XSLT-files mentioned. The workflow is illustrated below:

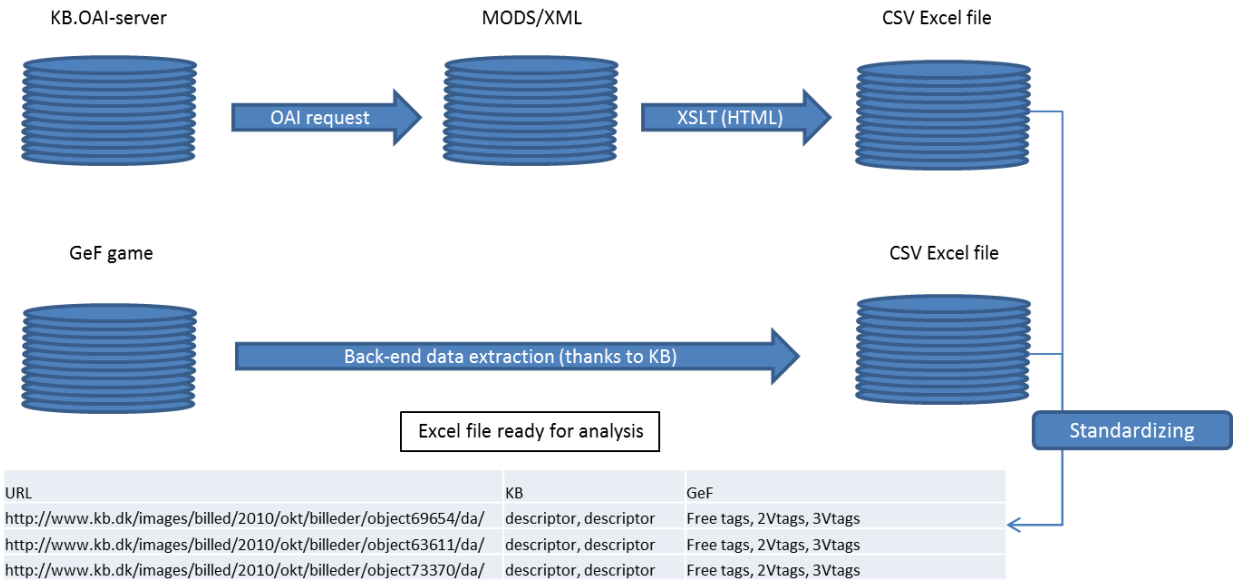

Figure 20 – Data collection workflow

[26] Castles and mansions
[27] Vacation and leisure time
[28] Christmas and Christmas customs
[29] Industry and crafts
[30] Churches and cemeteries

The final version, as seen in Appendix E included ObjectID, a URL to the digital object for speedy human validation, Free tags, 2Vtags, 3Vtags and the descriptors.

This resulted in three separate tag datasets for each type of tag:

Table 1

Total number of non-unique keywords				
	Free tags	2Vtags	3Vtags	Descriptors
	22787 (15525)	4743	2516	7306
Average	11	2,3	1,2	3,5

As the analysis calls for overlap, having the same term occurring multiple times in the Free tags could askew the measurements. It isn't likely that term (tag) weight will be used in a OPAC (and the current MODS as seen in appendix A isn't intended for it); should KB wish to use the Free tags in any way, they are not likely to include frequency in their metadata. This is why de-duplication of Free tags were performed on object level. The result is seen in the parenthesis in the leftmost column of Table X – expressing that the 2079 images was assigned a total of 15525 Free tags.

Table 2 shows the number of tags and descriptors assigned to the images. Only the Free tags occur in all 2079 pictures and the maximum number of Free tags added to an image is 30. The 19 images that don't have any descriptors assigned were indexed only with 'Denmark'.

The table itself provides a graphical overview of the validation process, as we see the number of Free tags assigned typically ranging from 5-15, while the 2Vtags and 3Vtags become increasingly 'top centered' i.e. that the typical image will have a much lower number of these kinds of tags. Of the 1517 images that had 3Vtags, i.e. a tag agreed upon by three separate players, more than half (53.1%) had only one 3Vtag.

Table 2

Number of terms assigned to images				
	Free tags	2Vtags	3Vtags	Descriptors
1	47	487	806	232
2	42	547	476	417
3	39	435	185	543
4	79	254	46	348
5	99	115	4	229
6	110	36	0	151
7	167	6	0	61
8	156	1	0	33
9	151	0	0	17
10	138	0	0	14
11	135	0	0	7
12	136	0	0	2
13	123	0	0	4
14	122	0	0	1
15	117	0	0	1
16	92	0	0	0
17	87	0	0	0
18	72	0	0	0
19	55	0	0	0
20	34	0	0	0
21	26	0	0	0
22	23	0	0	0
23	13	0	0	0
24	9	0	0	0
25	2	0	0	0
26	0	0	0	0
27	2	0	0	0
28	1	0	0	0
29	1	0	0	0
30	1	0	0	0
Occurs in	2079	1881	1517	2060

To facilitate term-category analysis on vocabulary level (RQ1), the unique values from each dataset were extracted by eliminating duplicates in excel. The resulting values are seen below in Table x.

Table 3

Number of unique keywords on vocabulary level				
	Free tags	2Vtags	3Vtags	Descriptors
Total	4121	1040	600	905

3.3 Research design

My research design is a combination of qualitative and quantitative methods. While RQ2 is purely quantitative, RQ1 & RQ3 both feature a mixture.

The initial creation of both term-categories (RQ1) and thesaurus-relations (RQ3) can be considered qualitative and relies of the researchers own interpretation of a) the relevant literature and b) the data at hand. Likewise, subjectivity is present in the actual coding, which features reading and classification. The final phase – counting and measuring the results – is quantitative.

Comparison of tags and descriptors on object level can be done on either syntactic or semantic level (Lu, Park, & Hu, 2010, s. 770). The former relates to whether tag and descriptor share the same word-stem (e.g. if the indexer and the player use the same word, but in plural and singular, respectively). The latter relates to whether player and indexer wanted to express the same basic notion (i.e. saw the same thing in the image) and used different, but related terms to express themselves.

While the syntactic measurement is a purely term-by-term undertaking, which can be done on the whole dataset, semantic overlap, in which the core meaning of a tag is held up against the descriptors, requires a higher degree of human interference (time) and is therefore only performed on a sample of the data.

The three methods are described and justified in separate sections below.

3.3.1 RQ1

RQ1: To what extent do tags (across all three validation thresholds) and descriptors fall within the same term-categories?

Essentially the data is comprised of four different vocabularies (datasets). Analysis is done on the unique terms from all four vocabularies (4121 Free tags, 1040 2Vtags, 600 3Vtags 907 descriptors).

Chapter 2.6.3 covers different pre-existing categories for describing image attributes and chapter 2.4.6 describes different tag categories. In relation tags, categories regarding personal information management such as ownership and re-retrieval were omitted in the initial categorization as the gameplay didn't lend itself to those types of tags.

The construction of categories is a form of content analysis, typically divided into the creation of a coding scheme and a definition of the recording units, then an assessment of the accuracy of the coding on a smaller sample, a revision and finally coding the entire text (Weber, 1990).

3.3.1.1 Theory generated codes

By using previous studies as a stepping stone, I utilize what Marshall & Rossman (2006) call "*theory generated code"* (p. 209); preliminary categories, informed by related literature. 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.

In chapter 3.6.3 a number of different frameworks for term-categories in the field of image indexing were presented. No consensus exists among the creators of these frameworks, although some ideas are ubiquitous: Object, event, location, time and interpretation. These informed my 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 *offness.*
- **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 term attempts 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 ones 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**
  Surprisingly, a small number of the Free tags contained obscene cusswords. While some of them were humorous, a few instances of racial/sexual slur made me want to isolate these as a separate category. Rather than the errors, this type of tag is probably what prevents direct seeding into catalogs.

The 'recording units mentioned' by Weber (1990) are the Free tags, 2Vtags, 3Vtags and Descriptors and the categorization process is depicted in figure 21:

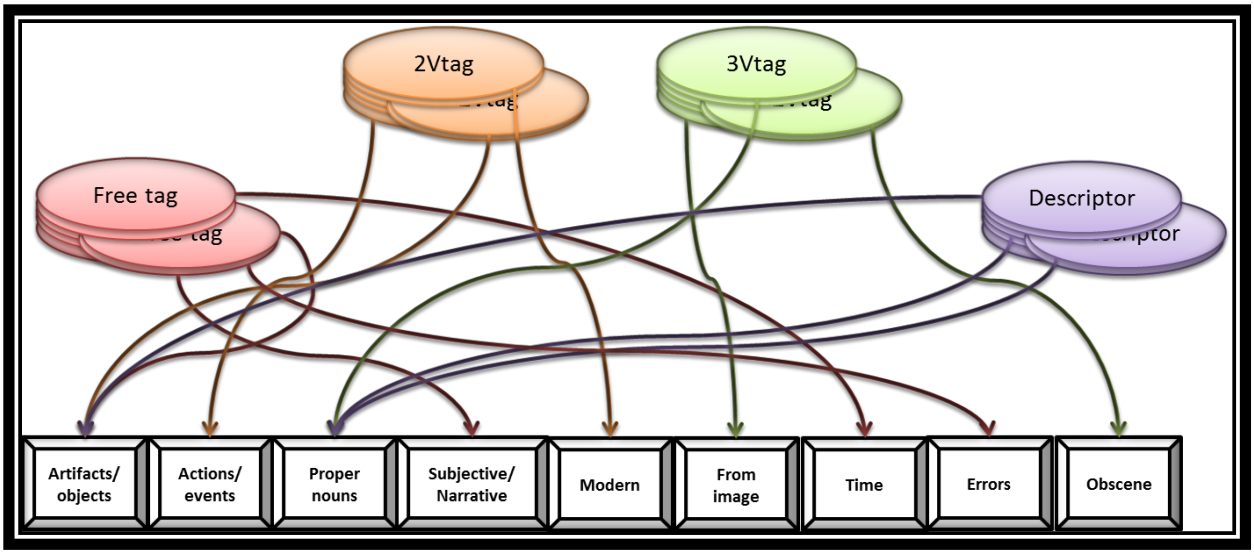

Figure 21 – Categorization process for RQ1

3.3.1.2 Compound terms

GeF 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' this option made sense, but also resulted in different kinds of compound tags, not referring to Proper Nouns. These compound tags all contained the character underscore and were initially isolated and then subjected to a refinement; four different subcategories of Compound terms were identified.

- **Two-term concepts** e.g. 'Flora_danica' or 'fishing net'. These are counted as 'Artifacts/objects'.
- **Refining tags or syncategorama** (Golder & Huberman, 2006) Tags which describe another tag in detail by serving as a qualifier i.e. adjective-noun pairs like 'old man' or 'short hair'. These were counted as 'Subjective/narrative'.
- **Title tags** Interesting 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 were counted as errors.

3.3.1.3 Limitations and discussion of method

As mentioned in the general limitations of the study (1.9), more researchers would probably have helped substantially, as both the creation of categories and the ensuing classification might have suffered from only having one viewpoint.

3.3.2 RQ2

RQ2: To what extent do taggers and indexers use the same terms to describe the same image?

This question can be seen as a way to discover 'inter-player/indexer' consistency and the analysis takes place at object level.

In order to determine the term similarity between the descriptors and the different groups of tags, a comparison on object level was made for each single image, pairing up potential tag-descriptor matches. As the data wasn't suited for this, each of the three sets of images (free, 2VTags, 3Vtags) was exported into XML, transformed in XSLT to create a set of rows like this:

object1 - tag1 - descriptor1
object1 - tag1 - descriptor2
object1 - tag1 - descriptor3
object1 – tag2 – descriptor1
object1 - tag2 – descriptor2
object1 - tag2 – descriptor3
object2 - tag1 - descriptor1
object2 - tag2 - descriptor1
object2 – tag3 – descriptor1

This allowed for comparison on object level, as each instance of a tag, was compared only to descriptors allocated to the same image. This is done to avoid mismatches between tags and descriptors of different images.

According to Peters, Schumann, Terliesner, & Stock (2011) a large portion of the research up to this point has been focused on vocabulary level comparison, i.e. looking at the folksonomy as a vocabulary and doing the analysis at that level. They suggest that analysis on object 'docsonomy' level (tags added to a single document) and the respective metadata is conceived as more *"valuable than overlap between folksonomy and the entire metadata collection"*(p.2).

The alternative to the overlap measured in this thesis would be a comparison on vocabulary-level, which would say less about actual image-indexing overlap and more about the vocabularies being used by indexers and taggers respectively (something already covered in RQ1).

The resulting of pairs with the 7306 descriptors was:

Free tags	**55634**
2Vtags	**16886**
3Vtags	**9061**

Each time a tag matches a descriptor, a score of 1 is assigned to the object. The similarity on object level can then be expressed by the Jaccard Index:

$$J(A,B) = |A \cap B| / |A \cup B|$$

Where |A ∩ B| is the number of matches, divided by |A ∪ B|, the total sum of tags and descriptors. Term overlap is simply where the normalized (capitalization removed) terms are determined to be the exact same string. Below are three tag-descriptor pairs, the first two of which are considered exact matches, but the third isn't.

København – København
Hund – Hund[31]
Pige - piger[32]

3.3.2.1 Exact and fuzzy matching

The third example Pige-piger can still be considered as a kind of term overlap, albeit not an exact one, which is why a more inclusive approach to measuring similarity should be taken. Rather than relying solely on an exact match on term level, fuzzy matching can give a clearer picture of syntactic overlap, as singular and plural variations of the same term are taken into account.

As we are dealing with a total of 81581 tag-descriptor pairs across all three validation thresholds, an automatic method is preferable to human computation.

One possible way to achieve this is to employ stemming software, which reduces words to their root and subsequently performing the analysis. The Snowball stemmer is an international project which originally was developed for English language but thanks to community support has been extended to 12 languages[33], amongst these Scandinavian. Using Snowball for this purpose does however require extensive technical knowledge, not available for this thesis. Another possible dataset for Danish language is available via the Danish Center for Language Technology, but this solution isn't free and was therefore foregone. Other, more approachable/affordable solutions exist, but most of the Open Source efforts in this field have been done for English language.

An alternative to stemming, which is entirely language independent, is suggested by (Mayfield & McNamee 2003, p.416) in which they describe N-gram tokenization.

3.3.2.2 N grams

The concept of n-grams was first discussed in 1951 by C.E. Shannon (Miller, Shen, Liu, & Nicholas, 2000, p. 4), famous for his mathematical theory of communication. N-grams are sequences of consecutive characters of the lengh *n* extracted from a document or corpus of text.

To generate the n-grams for a text, a window *n* characters in length is moved through the text, going forward one character at the time. At every position the sequence of characters in the window is recorded as an N-gram.

A list of the 5-grams in 'Sven Türck' are for example:
"_Sven", "Sven_", "ven_T", "en_Tü", "n_Tür", "_Türc", "Türck", "ürck_".

N-grams have been used in computational linguistics and information retrieval, notably in areas such as spelling-related applications, string searching, prediction and speech recognition (Miller, Shen, Liu,

[31] Dog/Dog
[32] Girl/Girls
[33] http://snowball.tartarus.org/

& Nicholas, 2000, p. 5). An example of useage is to enhance retrieval in OCR scanned texts. If the OCR software misreads a character, and "Sven Türck" becomes "Sven Türek" an exact match on the correct query, would yield unsatisfactory results. Matching on N-grams, rather than words - with a predetermined limitation of accepted deviance from exact matches - a user would still get the correct result, despite the OCR error.

By treating the morphological variations like spelling mistakes and using N-grams, it becomes possible to perform approximate (or fuzzy) string matching.

3.3.2.3 Levenshtein distance

The Levenshtein distance (LD) is another way to measure similarity between two strings of characters. The distance is defined as 'the minimal number of insertions, deletions or substitutions that are needed to transform one word into the other'.

- If tag is "Pige" and descriptor is "Pige", then LD(t,d) = 0, because no transformations are needed. The strings are identical to begin with.
- If tag is "Pige" and descriptor is "Piger", then LD(t,d)=1, because a single deletion (remove the letter "r") is sufficient to transform the string in the tag to the string in descriptor.

3.3.2.4 Excel algorithms

Fuzzy matching of terms is a method in which percentage values of term similarity are calculated by an algorithm. Then, a minimum value (depending on the matching algorithm used) is set and any tag-descriptor pair above that value is considered a fuzzy match. This is done to determine overlap where a tag is singular and the descriptor is plural. For instance, when a valid tag reads "Pige" and the descriptor reads "Piger".

A number of different algorithms[34] for determining term similarity are available as free code on the MREXCEL-website, each of them suited for a different purpose - none of them being measuring term similarity in Danish. And as each of them yielded unsatisfactory results on their own, it was decided to attempt to combine them. In order to test out the best configuration, each of the seven algorithms was tested alongside each other by human validation.

A sample of 1000 tag-descriptor pairs was extracted from the 2Vtags and each of the seven algorithms was tested to determine which one was best suited. The results were either very exclusive (i.e. too strict, resulting in matches not being made) or overly inclusive (i.e. too lenient, resulting in too many false positives).

By adding another algorithm from the same source, one that calculated the Levenshtein distance between two terms, and adding it to the equation, the approach yielded more satisfying results. Extensive testing was then done to determine the best combination of algorithms and in the end the best solution was deemed to be a combination of values produced by both the FuzzyPercent algorithm 1 and the Levenshtein distance calculated for each set.

Combining a minimum value of 0,8 for algorithm 1 with a maximum value of 3 for Levenshtein distance, creating a function in excel:

[34] Based on the ideas expressed in n-grams

=IF(AND(Algorithm1>0,8;Levenshtein distance<4);TRUE;FALSE)

The snippet below shows a small part of the spreadsheet used for testing the algorithms:

object_id	Vtag (2+)	Descriptor	Algorithm							Levenshtein distance (%)	Levenshtein distance	FuzzyMatch
			1	2	3	4	5	6	7			
61577	sommerhus	sommerhuse	1	1	1	0	0,08	0,19	0,24	89 %	1	1
61581	fanø	Fanø	1	1	1	1	1	1	1	75 %	1	1
61656	husmor	husmødre	0,83	0,64	0,7	0	0,05	0,08	0,12	50 %	3	1
76057	bro	børn	0,67	0,4	0,5	0	0,02	0,02	0,04	33 %	2	2
63652	hinke	hinkerude	1	1	1	0	0,05	0,09	0,13	20 %	4	2

The leftmost column shows the object number, the tag-descriptor pair is seen in the next two columns. Following that we see the seven different algorithms and the output they produce. The Levenshtein[35] distance in percentage and as a number is then displayed and in the farthest right column are the results of the function, where TRUE is 1 and FALSE is 2.

The chosen algorithm 1[36] performs the following: For each character in 'Tag', a search is performed on 'Descriptor'. The search is deemed successful if a character is found in 'Descriptor' within 3 characters of the current position.

A score is kept of matching characters which is returned as a percentage of the total possible score[37], essentially doing an n-gram calculation with a unigram and a proximity limit of 3.

In the third row from the table above, the maximum number of hits are 3 (number of characters in “Bro”) and the number of hits with a 3 character distance are 2, equaling a fuzzy match percentage of 67%.

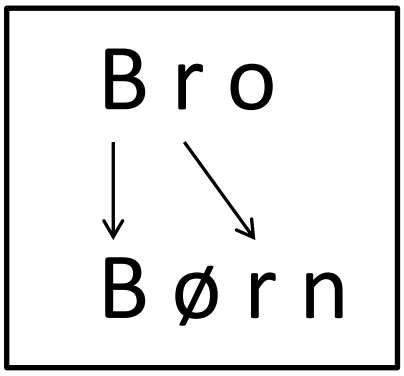

Figure 22 – Fuzzy match with algorithm 1

Essentially treating morphological variations of terms in the same way one would treat spelling errors, and by combining two well-known methods for approximate string comparison, we are able to determine the total number of fuzzy matches between the descriptors and each set of tags.

As tags in some rare cases can match up with more than one descriptor, e.g.:

Tag: Fisker - Descriptors: Fiskeri, Fiskere

Only one fuzzy match was allowed for each tag on object level.

3.3.2.3 Limitations and discussion of method

The example above highlights the weakness of this method – the false positives. Fisker (fisherman) and Fiskere (fishermen) should be fuzzy matched, but Fisker (fisherman) and Fiskeri (fishing) shouldn’t. As the algorithms only look at similarity form a mathematical point of view – rather than at word-stems, these errors are bound to occur.

35

36 http://www.mrexcel.com/forum/showthread.php?195635-Fuzzy-Matching-new-version-plus-explanation

37 http://www.mrexcel.com/forum/showthread.php?p=955137

3.3.3 RQ3

RQ3: To what extent do taggers and indexers use thesaurus-related terms to describe the same image?

RQ3 relates to semantic overlap; if tagger and indexer saw/described the same 'things', when they looked at the image.

Similarity and relatedness as semantic notions were discussed by Cattuto, Benz, Hotho, & Stumme (2008). They suggest two ways of defining similarity, either directly on the structure of the folksonomy itself (i.e. use statistical information to determine co-occurrence and distributions) or by mapping the tags to an existing ontology, thereby measure the relatedness of tags within a folksonomy to determine underlying data structures. This mapping of tags to an ontology can be taken one step further e.g. by mapping the relations *between* tags and descriptors to determine the 'semantic overlap' between a folksonomies and ontologies.

3.3.3.1 Thesaurus relations

While this mapping of relations could theoretically be based on highly expressive ontologies (chapter 2.3.1), previous studies have exclusively used thesauri for the mapping – a process known as thesaural term comparison.

The method was pioneered by (Voorbij, 1998) 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; title 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 (2005) to determine similarity between keywords assigned by authors, indexers and taggers respectively. Since then the Voorbij/Kipp approach has been used numerous times by the original authors (Kipp 2011; Voobij, 2012) and other researchers (Lykke et. al, 2011; Thomas, Caudle, & Schmitz, 2009; Wetterström, 2008). 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.

A thesaurus is an example of the ontolgies discussed earlier, and is a set of terms organised by their relationships to each other. Standard guidelines for constructing thesauri define three overarching types of relationships:

- Equivalence (equivalent terms U/UF)
- Hierarchical (e.g. broader/narrower terms: BT/NTs)
- Associative (Related Terms: RTs)

Figure 23 displays a small thesaurus with examples of these relations:

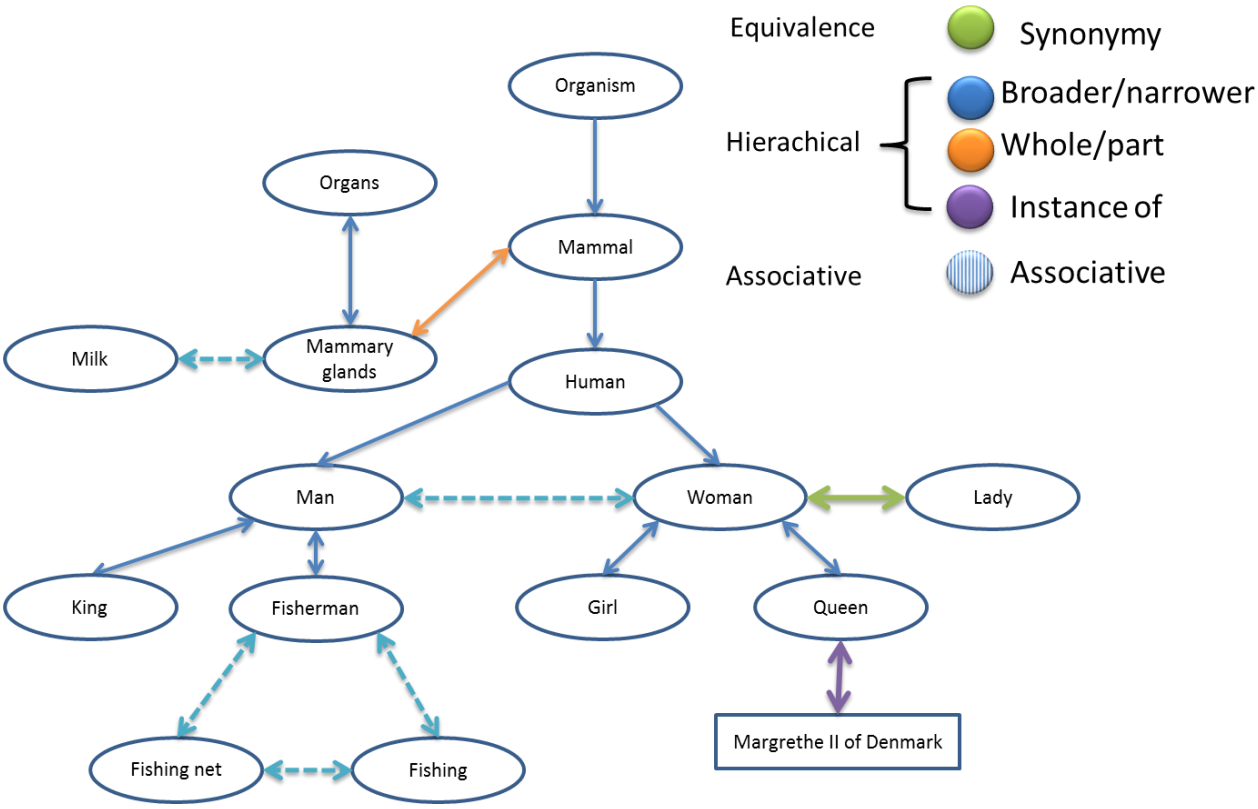

Figure 23 – Thesaurus with examples of the different types of relations

In addition to these, identical and non-related are also used as categories. In order to operationalize these different types, each of them are described and discussed below.

Equivalence relations

Basically, this is the same as the fuzzy syntactic overlap. If tag and descriptor share the same word stem, this relation is established. Not used in a normal thesaurus, as controlled vocabularies in their nature will have rules dictating a how a word is noted (usually plural, like it is the case for the descriptors in the Sven Türck collection).

Equivalence relations

Equivalence is established if tag and descriptor are two different terms with the same meaning (i.e. synonyms). In a thesaurus this relation is typically denoted by Use (U) and Use-For (UF). This is done when an authorized term is to be used instead of an unauthorized one e.g. when a law has a popular name, but also has a more formal number in a controlled vocabulary.

'LOV nr 907 af 15/10/1996' **Use For** 'Rockerloven'

And the reciprocal:

'Rockerloven" **Use** "LOV nr 907 af 15/10/1996'

When deciding equivalence, however, one has to take the two distinct problems of natural language into account (Svenonius, 2000, p. 148). First there is the problem of homographs, which occurs when a single term has two or more meanings:

The Danish word for lady, “dame” has two distinct meanings: the first being “lady” and the second being the word for a playing card, which incidentally is called “queen” in English. In KOS these two meanings are qualified somehow (e.g. by a disambiguating parenthetical notation) to distinguish between them.

The second issue is that of polysemes - identical words that share meaning, either psychological or etymological, but cannot be said to express the exact same concept. In the Danish dictionary, the aforementioned “dame” has three distinct meanings, each describing a slightly different meaning of the word[38].

For our purposes, human validation by looking at each picture isn’t feasible, so when the term “dame” appears as a tag, and the Danish word for woman “kvinde” is used as a descriptor, we assume that no playing cards are depicted and equivalence is established.

Hierarchical relations

“*Our brains are hardwired to perceive hierarchical relationships, and, consequently, the only way to comprehend a knowledge domain is through the structure they provide*” (Simon, 1962 as cited by Svenonius, 2000).

In a typical public library in Denmark this relationship becomes clear the moment one searches for a book amongst the shelves; the DK5 classification borrows heavily from DDC, and consists of 10 superclasses and as hundred subclasses – and so on. As one walks along the shelves, the numbers on the back of the books become longer, as the (main) subject within becomes more and more specific.

5X Natural sciences

51 Mathematics

51.3 Differential equations

The hierachical relation links broader and narrower terms. Three different kinds of hierachical relations can be coded by using an ‘is a’-test.

- is a type of, e.g. a cow is a type of mammal
- is a part of, e.g. a finger is a part of a hand
- is an instance of, e.g. Haley's comet is an instance of a comet

Broader/narrower

The first one, the *type of*, can also be classified as genus-species or inclusion relationship, with the inheritance property that ‘what is true for mammal, must also be true for cow’ i.e. a class inherits all properties from its parent-class.

Whole/part

The *whole-part* relationship is similar to broader and narrower terms. If the tag “finger” appears and the term “hand” is found in the corresponding set of descriptors, the relation noted is part-whole.

[38] http://ordnet.dk/ddo/ordbog?query=dame&tab=for

Aside from the physical component part relationship, topic-subtopic, and region-subregion relationship are also treated as this type of relations.

In case of two proper nouns e.g. the tag 'Vesterbro' (a borough in Copenhagen) and the descriptor 'Copenhagen', there is also a *part-whole* relation – in which case they are considered hierarchically related terms.

Instance of

The last 'is a' test, the *instance of*, is useful when we are coding the tags or descriptors, where one of them is a proper noun. If the tag describes the broader term of the Proper noun, such as "Castle" – "Kronborg" there is a semantic relation as the tags denote the type of things described. So a broader type of hierarchical relation is established if the proper noun in the descriptor can be said to be an instance of the tags. The reverse example "The Brave Soldier" – "memorial" where the tag specifies exactly which memorial is depicted, will accordingly be coded as a narrower type of hierarchical relation. These relations are named *Tag-literal* and *Literal-descriptor*.

Associative relations

Associative relations are by far the least consistently applied thesaurus relation.

Svenonius notes how guidelines for controlled vocabularies usually define the associative relationships broadly and vaguely (i.e. inclusive), to simply encompass all semantic relations, except for those of equivalence and hierarchy and points out a general lack of rigor when determining the exact conditions under which two terms are considered to be related. She provides an example from the early days of LCSH in which two terms were allowed to be considered related, if they occurred in the same monograph (Svenonius, 2000, pp. 160-161) – a rather extreme version of the syntagmatic relation, which would result in "romance" being linked to "grass" and "scarf" to "landscape"[39]. The only formal mathematical property underlying the notion of relatedness is that of symmetry: if "term 1" is related to "term 2", then "term 2" is related to "term 1".

In their article on tag similarity, Cattuto, Benz, Hotho, & Stumme (2008) even notes that: "*In most studies, the selected measures of relatedness seem to have been chosen in a rather ad-hoc fashion*" (p.616) which hardly seems surprising, when keeping Svenonius' musings on the relative nature of 'associative relatedness' in mind.

Bechhofer & Goble (2001) list various categories of associative relations found in thesauri, including examples such as: an occupation and the person in that occupation, e.g., "fishing" and "fisherman"; a thing or action and its counteragent, e.g. "pests" and "pesticides"; and an action and its product, e.g. "harvesting" and "harvest", but notes that these are only examples of the categories found in the wild. An example of another associative relation than the ones mentioned above found while coding the data is:

Definition-dependent: when a tag is mentioned in the dictionary definition of a descriptor (or vice versa).

[39] Example from Sven Türck picture with object_id 74121

E.g. when the tag reads 'paddle' and the descriptor read 'kayak'. These aren't linked as nodes in the reference standard, but the definition of the word 'paddle' is that it is a type of oar 'used for rowing a kayak' making for an intuitive way to establish an associative relation.

3.3.3.2 Reference standard

The other studies in which term comparisons have been done, usually use the 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 LCSH or MEDLINE, the relation is established. But a looser interpretation of 'associative' has been taken by several researchers (Thomas, Caudle, & Schmitz,2009; Lykke, Hoj, Madsen, Golub, & Tudhope, 2012).

As mentioned (chapter 1.3) the descriptors are comprised of categories, locations and people. While the categories have a hierarchical structure, it is only made for the specific image database and isn't rich enough to be used as a reference standard. The tag "puppy" is for example clearly a narrower term of "dog", but as it doesn't appear in the hierarchy of categories used by KB, this relation wouldn't be established by relying on the pre-existing categories.

Without a reference standard, 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.

AndreOrd provides a visualization tool of the Danish language in which the relations between terms are color-coded. An example for the term bicycle (cykel) is given below in figure 24.:

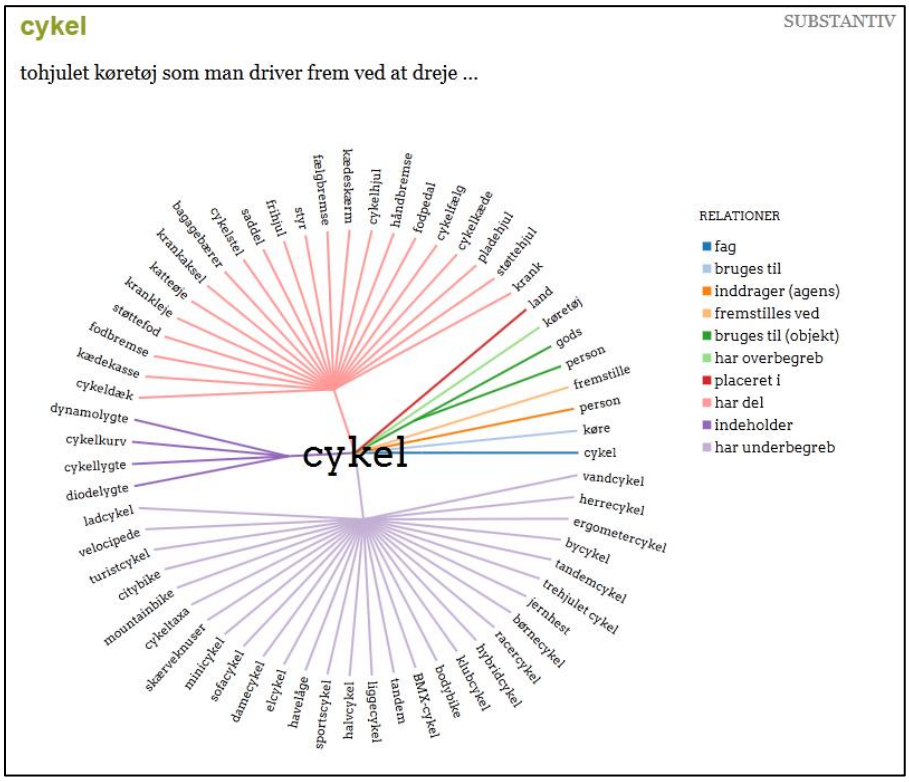

Figure 24 Screenshot of relations with the term 'cykel' from andreord.dk

AndreOrd also has a number of other types of relations and, in the analysis, all of these other relations are classified as associative.

Using an external set of data might seem problematic at face value as the chaotic tags might not exist in such a resource – but a study has shown that roughly 61% of the 10.000 most frequent unique tags in delicious are found in WordNet (Cattuto, Benz, Hotho, & Stumme, 2008, s. 10) and AndreOrd turned out to be very reliable.

The depiction of 'cykel' in AndreOrd hints at the same network structure also seen in my homemade thesaurus example (Figure 22) the concepts are basically a network of interconnected nodes.

3.3.3.4 The 'small world' of thesaurus relations

A challenge when looking for thesaurus relations is hinted by the network structure it represents. For example, when faced with the tag "window" and the descriptor "house" we face an interesting dilemma, which re-occurs often when attempting to categorize relations.

Without seeing the image, we imagine a house and – like most houses – this house has windows. When querying for "window" in AndreOrd, there is no direct relation between the two terms.

If we add the term 'building' to the equation - which has a whole-part relation to 'window' and a broader/narrower relation to 'house', the relationship is established. But in doing so, we define our relations by two degrees of separation (figure 23).

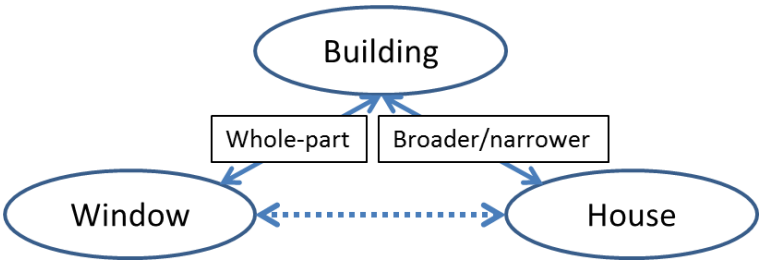

Figure 25 - Concept of distance in a thesaurus

This idea of distances is called the 'small world phenomenon' and it appears in any kind of network. By taking a mathematical approach to the 'distances' between two words in a thesaurus, they found that the average 'distance' between two completely random terms in the English language is 3,05 (Motter, Moura, Lai, & Dasgupta, 2002). Another example from the other spectrum of our knowledge representation universe is given by Schmitz, et al. (2007) describing that every user, tag or resource in the folksonomy delicious can be reached by pivot browsing from any other user, tag or resource by an average of 3,5 mouse-clicks.

When, for instance, determining hierarchical distances - taken to the extreme - all 66.300 terms in AndreOrd are related 'upwards' to the super-category "object" by very few nodes. In order to ensure rigor in the analysis, I chose a maximum degree of separation of x=1 for all relations in AndreOrd. For my thesaurus example in figure 23 this means that a 'queen' is a 'woman' but not a 'human' or a 'mammal'.

The exceptions to this are the Tag-Literal and Literal-Descriptor relations; proper nouns do not appear in AndreOrd and the connection has to be made manually and is therefore more subjective in nature.

3.3.3.5 Coding

The coding is done exclusively, meaning that for each image, every tag is compared to all the descriptors and a single match type is determined (Lykke et. al, 2011, p.44) via a prioritized list i.e. that certain relations are considered 'stronger' than others. In some cases a tag might have a relation to more than one descriptor.

- Tag: Fisher[40]
- Descriptors: Fishing, Fish, Fisherman[41]

In which case the Same-relationship is chosen over the *associative*. The order of importance was inspired by (Lykke et. al, 2011) with same/synonym being the strongest, followed by narrower (superceeding broader) and ending with the loosely defined associative relation. For the coding the relations 5-8 were added.

1. Same (similar to fuzzy syntactic match)
2. Equivalence (synonym)
3. Narrower Term
4. Broader Term
5. Part-Whole
6. Whole-Part
7. Literal-descriptor
8. Tag-literal
9. Associative

In the end the analysis was done by counting the number of relations, following the workflow seen in figure 26:

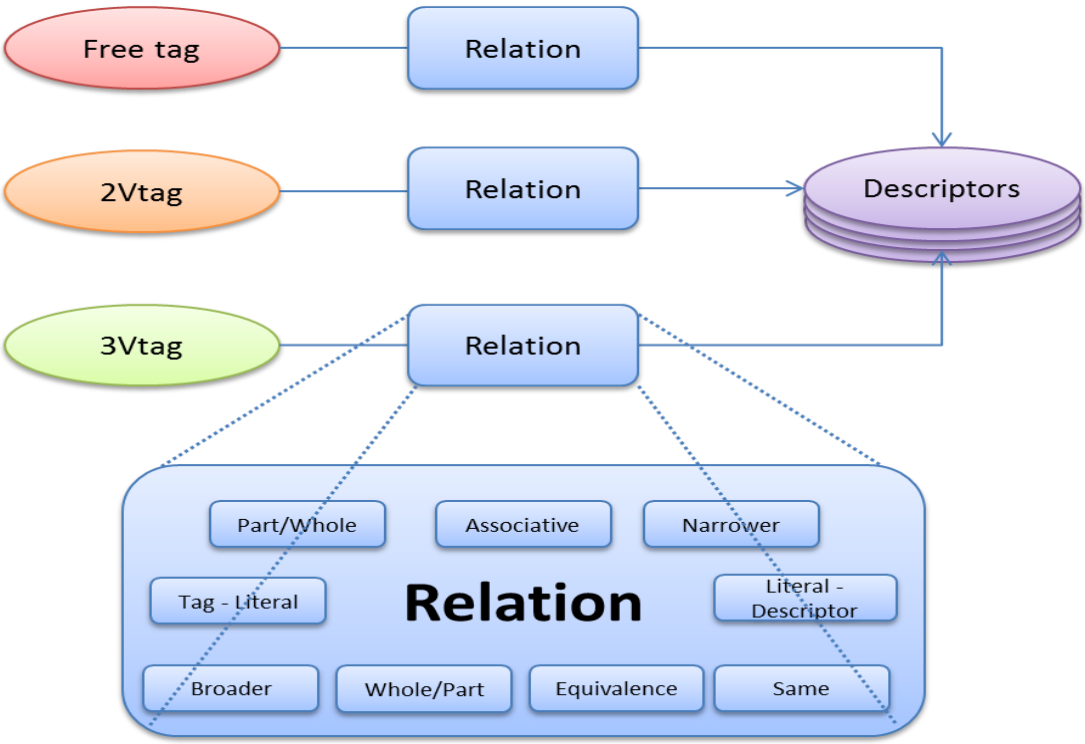

Figure 26 Thesaurus relation analysis

[40] Fisker
[41] Fiskeri, fisk, fisker

3.3.3.6 Limitations and discussion of method

Even though one type of relation takes precedence over another, there is no meaningful way to measure the distance between them, meaning that this type of study is taking place at what Wetterstrom (2008) refers to as a '*nominal level of measurement*'(p.295). Attempts have been done to quantify the thesaurus distances in a vector space. Medelyan & Witten (2006)did a study on inter-indexer inconsistency using thesaural term comparison. As tags and descriptors simply can be seen as the output of two independent indexers, methods from these types of studies should be considered . The researchers assigned the weight 65% to synonyms, 20% to related and 15% to hierachical terms, but this can only be done when working with a strict reference standard with clear cut definitions of related terms and as described, this isn't the case for the Sven Türck collection.

In a recent tag-descriptor study, Voorbij (2012) states how tags are "*classified as (nearly) exact to, synonym of, broader than, related to, narrower than or not represented in the subject heading or any other keyword of the record. This was done without the help of a thesaurus and without examining the books*" (p.214)[42] highlighting a somewhat loose approach to thesaural relations.

In his investigation into complementarity of tags and LCSH in a New Zealand context, Wetterstrom (2008) admits a certain degree of subjectivity involved in the analysis, writing how: "*The division of tags into groups and categories was not always clear in that many tags could qualify for more than one group or category*" (p.297) and Kipp (2005) in one of the studies pioneering the method, can also be said to be guilty of this somewhat lax approach to relatedness. She uses seven categories, and almost half (44.5%) of the established relationships fell into the 6th category – 'related but with some ambiguity in the relationship' of which common relationships included: "*The relationship between an object and its field of study, the relationship between two fields of study which examine different aspects of the same phenomenon, and the use of a methodology or form of inquiry in a new environment*" (p.9) but were not limited to those.
Another hierachical relationship, the perspective relationship, allows for a less strict interpretation of this type of relation and lets the indexer express a point of view e.g. a rat being an agricultural pest, rather than just a type of rodent. While limiting, for the purpose of this analysis, perspective hierachical relationships are omitted, as they are 'context-dependent and not defintionally true' (Svenonius, 2000, p.164).

The usage of AndreOrd in combination with this strict adherence to 'distance' is an important part of the research design of this study, as it minimizes the impact of the individual researcher, i.e. ensuring inter-subjectivity by using an objective standard.

3.4 Chapter summary

In this chapter I have accounted for the data collection process behind obtaining the data and described and discussed the different methods I used to analyze the data in order to answer the research questions.

[42] Underline added

Chapter 4 Findings and discussion

4.1 Chapter overview

In this chapter I will present the findings related to each of the three research questions.

For each question there is an initial presentation of the findings in tables, followed by a discussion along with examples of tags, descriptors and the images to which they were added[43], along with visualizations of the data in the form of graphs and Venn diagrams.

4.2 RQ1 findings and discussion

RQ1: To what extent do tags (across all three validation thresholds) and descriptors fall within the same term-categories?

Initially the 4121 unique Free tags and 905 unique Descriptors were classified according the 9 term-categories. The 1040 unique 2Vtags and 600 unique 3Vtags were classified by cross-referencing with the classification done for the Free tags as they consisted of the same words. This was done partly to save time and partly to ensure consistency i.e. making sure that the same word was classified in the same way across all three levels of validation.

The 218 compound terms found in the unique Free tags were treated separately and divided into their assigned categories, as shown in figure 27:

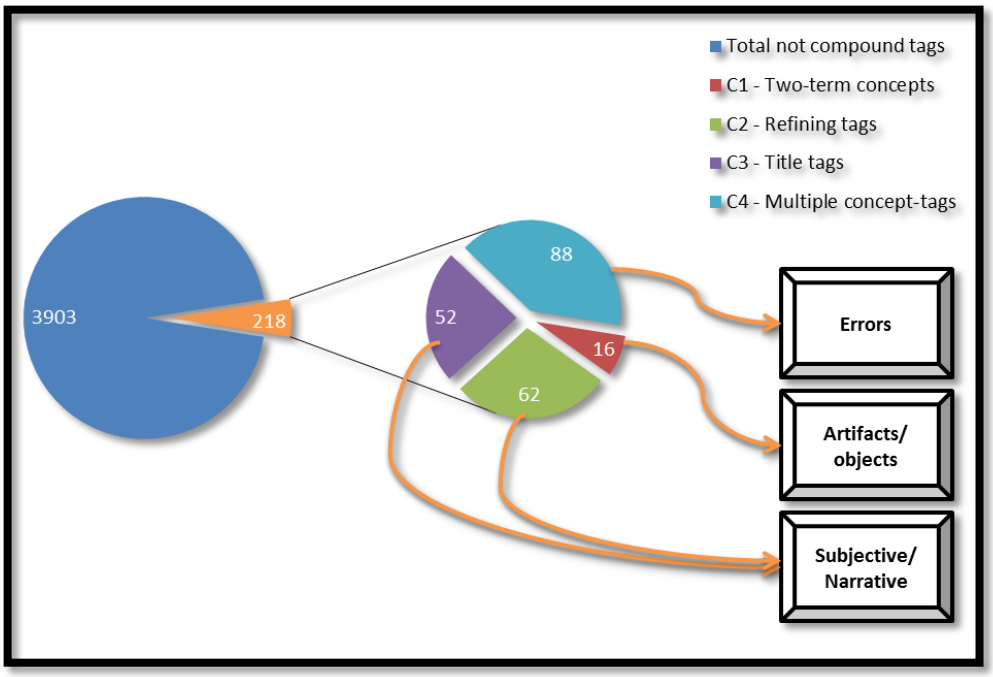

Figure 27 – Analysis of compound terms in the Free tags

[43] Examples will appear in English, with the original Danish terms shown in parenthesis in the discussion.

The final distribution is seen in table 4.

Table 4
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

This expresses the vernacular vocabularies of each of the four datasets i.e. the dictionary of which each dataset is comprised. It clarifies how many of each unique term belongs to the same category, but it does not say anything about the actual distribution. For instance are there 34 unique terms from the Free tags belonging to the 'Time' category – but without knowing the actual occurrences of them i.e. the non-unique distribution, we only see half the picture.

The categorization was therefore used to calculate the distribution for non-unique tags (table 5).

Table 5

Term-category distribution among non-unique terms

	Free tags (*n*=2079)[44]			2Vtags (*n*=1881)		
Category	Frequency (%)	Cumulative %	*M (SD)*	Frequency (%)	Cumulative %	*M (SD)*
Artifacts/objects	12271 (79%)	79,04	5,9 (2,93)	4185 (88,24%)	88,24	2,22 (1,3)
Actions/events	831 (5,4%)	84,39	0,4 (0,86)	185 (3,9%)	92,14	0,1 (0,35)
Proper nouns	909 (5,9%)	90,24	0,44 (0,82)	288 (6,07%)	98,22	0,15 (0,39)
Subjective/narrative	583 (3,8%)	93,99	0,28 (0,61)	39 (0,82%)	99,04	0,02 (0,15)
Modern	56 (0,4%)	94,35	0,03 (0,17)	3 (0,06%)	99,1	0 (0,04)
From image	13 (0,1%)	94,43	0 (0,06)	1 (0,02%)	99,12	0 (0,02)
Time	71 (0,5%)	94,9	0,03 (0,19)	12 (0,25%)	99,37	0,01 (0,08)
Errors	762 (4,9%)	99,8	0,37 (0,64)	30 (0,63%)	100	0,02 (0,13)
Obscene	30 (0,2%)	100	0,01 (0,12)	0 (0%)	100	0 (0)
Total	15525 (100%)		*7,46 (6,4)*	4743 (100%)		*2,52 (1,33)*

	3Vtags (*n*=1517)			Descriptors (*n*=2062)		
Category	Frequency (%)	Cumulative %	*M (SD)*	Frequency (%)	Cumulative %	*M (SD)*
Artifacts/objects	2245 (89,2%)	89,2	1,48 (0,86)	4479 (61,3%)	61,31	2,17 (1,77)
Actions/events	97 (3,9%)	93,06	0,06 (0,28)	590 (8,1%)	69,39	0,29 (0,59)
Proper nouns	149 (5,9%)	99	0,1 (0,3)	2062 (28,2%)	97,6	1 (1)
Subjective/narrative	8 (0,3%)	99,32	0 (0,06)	117 (1,6%)	99,2	0,06 (0,23)
Modern	1 (0%)	99,36	0 (0,03)	0 (0%)	99,2	0 (0)
From image	0 (0%)	99,36	0 (0)	0 (0%)	99,2	0 (0)
Time	3 (0,1%)	99,48	0 (0,04)	37 (0,5%)	99,71	0,02 (0,13)
Errors	13 (0,5%)	100	0,01 (0,09)	21 (0,3%)	100	0,01 (0,1)
Obscene	0 (0%)	100	0 (0)	0 (0%)	100	0 (0)
Total	2516 (100%)		*1,66 (0,87)*	7306 (100%)		*3,54 (1,96)*

In the remainder of the section an example of an image from each term-category will be displayed and their frequency-distribution discussed.

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

Figure 28 – Artifacts/objects & Actions/events

1: Artifacts/object

We see a number of typical terms describing what is 'in' the picture: Ball (bold), flowers (blomster), grass (græs), girls (piger) and sky (himmel). These types of terms were especially dominant for 2Vtags and 3Vtags constituting 79,7% and 84,4% of the vocabulary and 88,24% and 89,2% of the non-unique terms. Only 56,9% of the unique Free tags are from the same category and when looking at the non-unique Free tags, 79% are 'Artifacts/object', which shows that this category has many re-occurrences of terms and therefore also a high validation rate. As the terms in this category typically describe things from Panofky's Pre-iconographic level, we can infer that simple terms at that level of description are more likely to be validated in this type of game.

For Descriptors around half the unique terms (51,8%) belong to 'Artifact/objects' and slightly more 61.31% for the non-unique terms, which is higher than the distribution of 22.51% found for 'Objects' in St. Andrews Library Photographic Archive (Rorissa, 2010).

2: Actions/events

Two girls on a beach, Jumping (hoppe), Running (løbe) and Skipping (springe).

The terms describing 'Actions/events' were a lot rarer than 'Artifacts/object' for all four datasets - but there was a fairly high percentage of these terms in all three kinds of tags, but a relatively low frequency in comparison, suggesting that a unique 'Actions/events' term doesn't tend to be repeated a lot for tags.

The opposite is true for descriptors, where the value for unique terms in this category is only 3,4% compared to the 8,1% for non-unique terms; there are only 31 different 'Actions/events' terms, but these 31 terms are being used 590 times (or 19 times on average). A closer investigation shows that 212 of the 590 occurrences in this category are only two unique terms: 'vacation' (172 occurrences) and 'Christmas' (40 occurrences).

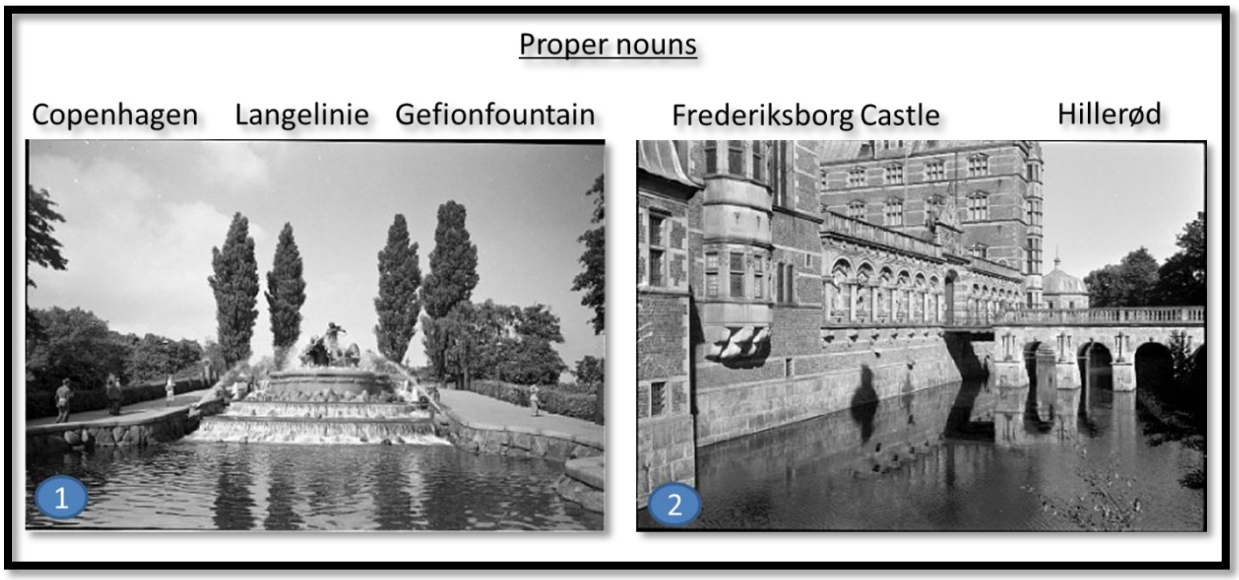

Figure 29 - Proper nouns

1 & 2: Both pictures are of Danish landmarks with both the names of the landmarks and their geographic location[45]. In looking at non-unique terms for this category, it is found the almost a third (30,3%) of all the occurrences in the Free tags are the made up by the same 11 locations. The same is seen in 3Vtags, where only ten unique proper nouns occur more than 2 times, and all but one (Dannebrog – the national flag of Denmark), are famous landmarks:

Table 6
Non-unique 3Vtags with more than two occurrences in 'Proper noun' term-category

Non-unique terms	Frequency
Tivoli	30
Rådhuspladsen	18
København (Copenhagen)	12
Amalienborg	11
Kronborg	8
Christiansborg	7
Børsen	6
Dannebrog	6
Nyhavn	6
Frihedsstøtten	5
Total	109

[45] Location was found to be the predominant form of 'Proper noun' across all four datasets.

These 109 non-unique term make up 73,1% of all the 3Vtags in that category.

Unlike an indexing situation, the game does not give the player time to consider or check with external sources, so if the location isn't instantly recognizable to the player, this type of term will not be added in a game. Whether a 'Proper noun' is added (and validated), depends on how 'well-known' the motif in the image is.

For descriptors, 'Proper noun' was the second largest term-category for both unique (42,2%) and non-unique (28,2%) descriptors. This reflects an indexing policy with focuses on this type of descriptor and corresponds with Chens (2001) study of art history queries, where 'unique locations' were the predominant type of query. The distribution 28,2% for non-unique terms is significantly lower than the 43.35% distribution for non-unique terms categorized as 'Locations' in St. Andrews Library Photographic Archive (Rorissa, 2010).

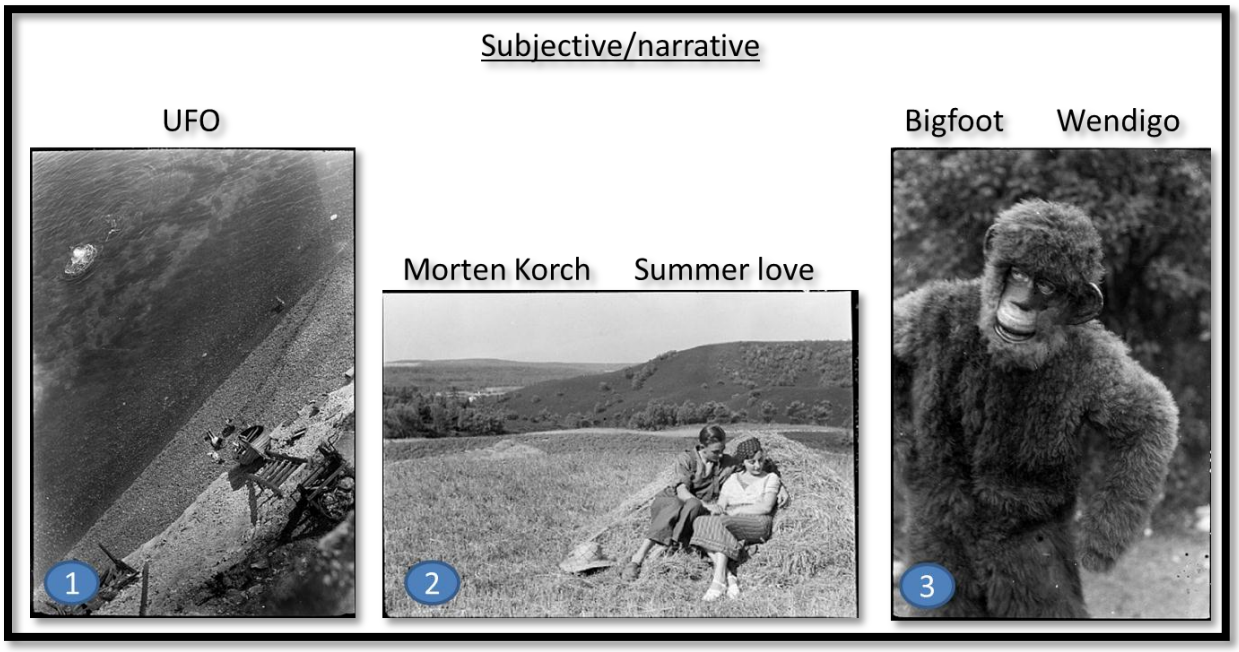

Figure 30 – Subjective/narrative

1: A coastline seen from above. A player probably saw a UFO in the splash of water in the upper left corner.

2: 'Morten Korch' refers to a famous Danish author from the same period as Sven Türck, whose books are about the same idyllic version of Denmark as Türck's photographs. Summer love (sommerforelskelse) shows the mood in the picture.

3: Two very narrative terms to describe someone in a monkey costume, each referring to a mythical ape-like creature, native to North America.

The three first term-categories 'Artifacts/object', 'Actions/events' and 'Proper nouns' were to various degrees represented across all four datasets. For the 'Subjective/narrative' term-category, a drastic drop happens for the unique terms from Free tags (9,2%) to 2Vtags (2%) to 3Vtags (0,5%). The same is found for non-unique terms with 583 (3,8%) for Free tags and only 8 (0,3%) 3Vtags belonging to this category.

For descriptors we see the same pattern as we saw for the 'Actions/event'; a small number of terms are repeated numerous times. The 'Subjective/narrative' descriptor 'Youth' alone makes up 68,3% of all occurrences of the term-category (80 out of 117).

Figure 31 – Subjective/narrative compound terms

1: View though opening in forest (udsigt gennem skovåbning) which could serve as a title for the picture.

2: Happy men (glade mænd) is a typical example of the coupling of an adjective and a noun in the free tags. Other examples are 'Tall hats', 'Resting people', 'Small children' and 'Short hair'.

The 114 compound 'Title tags' and 'Refining tags' (see figure 27) were never validated and can only be found among the Free Tags. Not surprising, considering their idiosyncratic nature.

Three of these types of tags were found in the descriptors from the 'Subjective/narrative' category:

- Country idyllic (landlig idyl)
- Small boy at the water's edge (lille dreng i vandkanten)
- Mothers with kids (mødre med børn)

This is where the Free tags differed considerably from the other three datasets and it is clear that the uncontrolled tags are much more likely to express interpretations of the image, i.e. aboutness.

As the 'Subjective'narrative' category encompassed all adjectives, different results might have come from a collection of color photography as more colors might have been described.

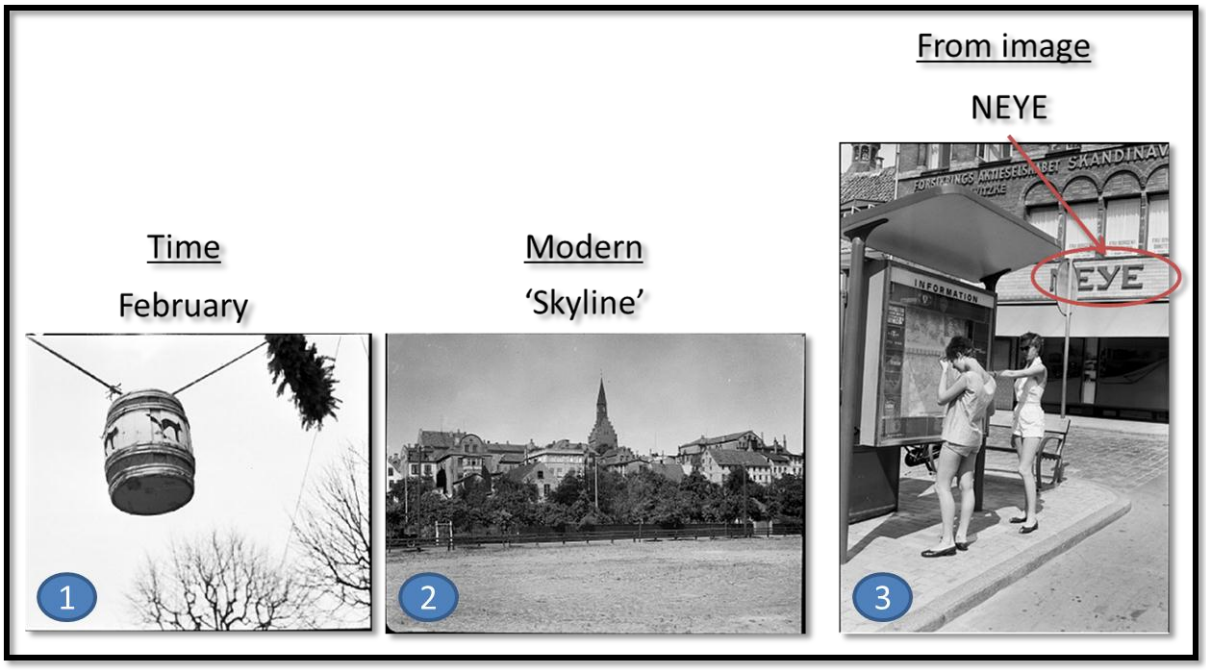

Figure 32 – Time, Modern and From image

1: Depicts a special and recognizable type of decorated barrel used in the Danish carnival (Fastelavn). The tag is a very good example of clever tagging as 'Fastelavn' usually takes place in February, depending on the liturgical year.

2: The tag directly read 'skyline' an example of an English word that recently made it into the Danish Dictionary, but probably wouldn't normally be associated with this type of old photography.

3: An example of a 'From image' tag that has no meaning without the image. NEYE refers to a specific shop in central Copenhagen.

These term-categories were almost exclusively found in the Free tags and even here they were infrequent. The lack of 'Time' in descriptors might be due to the fact that the entire collection has a sort of 'timestamp' attached to it; Sven Türck's pictures, on a general level, depict a specific Danish era, and the time is therefore implied. If the photographs did not contain specific dates, it is unlikely that a professional indexer would just guess, as happened with player and the decorated barrel. The Danish Carnival does after all, sometimes take place in early March, and an indexer might not run the risk of an error based on a qualified guess.

Figure 33 – Screenshot with spelling error from KBs website.

One of the advantages of validating tags, aside from players agreeing on the conceptual level is the automatic spell-checking mechanism imbedded in the validation process. Even though players tag completely uncontrolled, the odds that two players make the same mistake is low.

For the Free tags 575 (14%) of the unique terms were errors, either spelling mistakes, typing mistakes or 'Multiple concept-tags'. The validation effectively eliminates almost all errors and we see a lower error-rate for the unique terms in 2Vtags 9 (0,9%) than for the descriptors 12 (1,3%). In the example above one of these errors is taken from the website.

Obscene

A total of 30 non-unique and 18 unique tags in the pool of Free tags represented 'Subjective/narrative' tags which could be seen as offensive. These ranged from fairly lighthearted humoristic names for body parts to a few cases of sexual and racial slur[46]. While much rarer than errors, this type of tag represents is a stronger argument for not seeding Free tags directly into the catalog without some sort of manual control. Just one single case of a racist tag in the catalog would reflect poorly on the library.

[46] Examples are omitted.

In general a few unique tags/descriptors formed the majority of the four datasets:

Table 7

Number of unique keywords constituting half the dataset				
	Free tags	2Vtags	3Vtags	Descriptors
	212	75	52	58

Each of the four datasets follows a standard power law distribution, with long trunk and a long tail. Figure x shows these distributions with unique terms on the x-axis and number of occurrences on the y-axis.

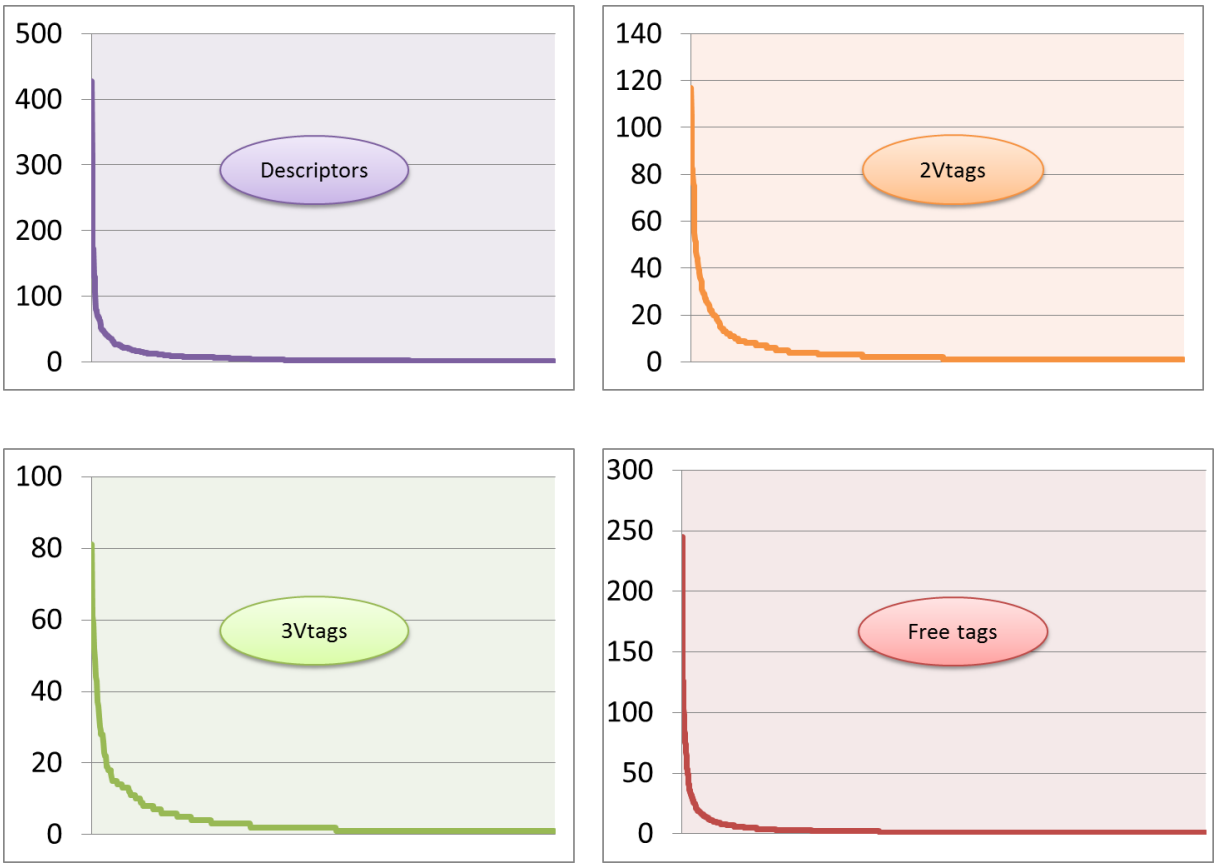

Figure 34 - Power law distribution in each of the datasets

The GeF validation threshold was set so that only tags that were repeated three or more times had any chance of ever being validated (the 3Vtags). All the Free tags that occurred just one or two times can be said to form a sort of long tail, where the majority of the 4180 unique terms from the Free tags is to be found.

Table 8
Unique term-category distribution of Free Tags occurring less than three times

Term category	Frequency	% of categories
Artifacts/objects	1552	66,2 %
Actions/events	310	79,1 %
Proper nouns	235	74,4 %
Subjective/narrative	340	89,5 %
Modern	49	98,0 %
From image	11	100,0 %
Time	27	79,4 %
Errors	564	98,1 %
Obscene	16	88,9 %
Total	3104	100 %

More than 75% of the vocabulary in the Free tags belong to this 'long tail'. Figure 35 and table 8 displays how the long tail is where 80% of 'Actions/event' 90% of the 'Subjective/narrative' tags resides. This is also where 98,1% of the 'Errors' are found along with *all* of the 'obscene' tags.

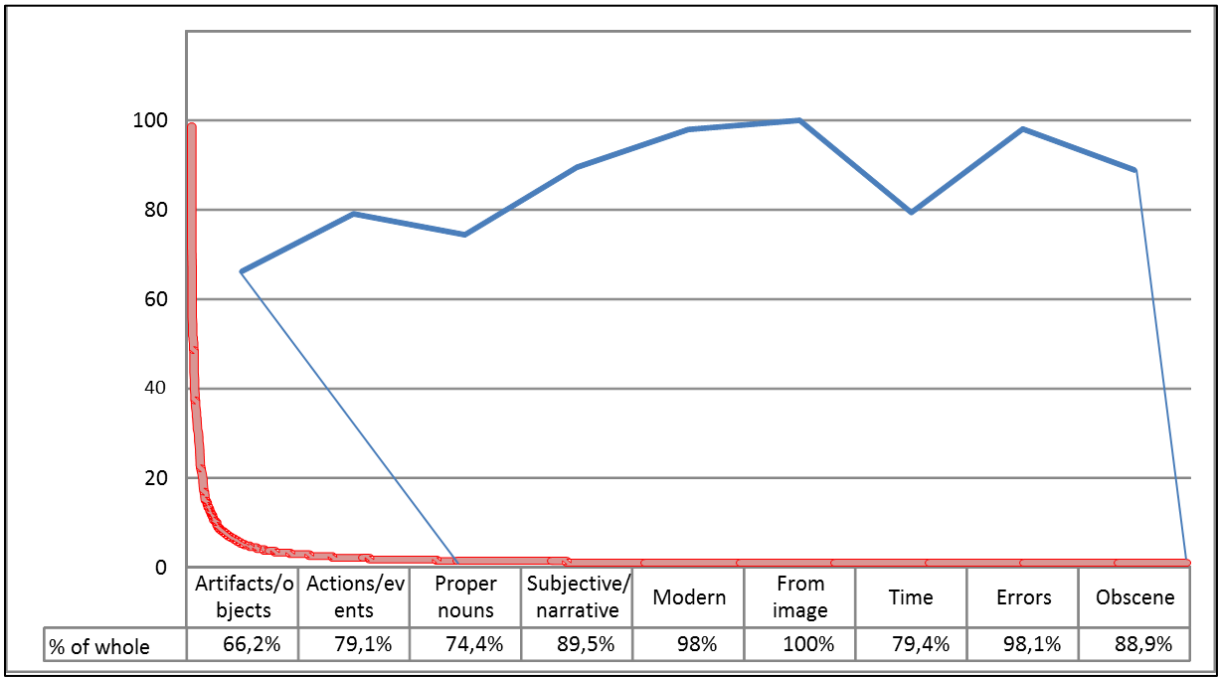

Figure 35 – Long tail distribution of unique Free tags among term-categories

These long tail tags are unlikely to be validated as 2Vtags (and would never occur in the 'official' 3Vtags), but they do represent both the semantic richness of the folksonomy and the flipside of the coin very well.

On one hand, the validation process removes errors and even potentially harmful obscene tags, but potentially valuable descriptions might also be lost. This does, for example, hold true for the compound terms that were categorized as 'Title tags' (figure 31).

The descriptors belonged mainly to 'Artifact/objects' and 'Proper nouns' with a smaller subset from the 'Actions/events' term-category. Both 2Vtags and 3Vtags also belonged almost exclusively to these three term-categories, but had a much larger proportion of 'Artifacts/objects' and a much lower proportion of 'Proper nouns'.

Even though the coding wasn't done for word-categories (nouns, adjectives, verbs) the 'Artifact/objects' are nouns and the analysis confirms earlier findings that most tags are nouns (Heckner, Mühlbacher, & Wolff, 2008; Guy & Tonkin, 2006).

The validation process in this type of human computation game is therefore shown to favor 'Artifacts/objects'. This term-category contains the terms that explain what is seen *in* the image, rather than what it is *about* and we can conclude therefore that subjective terms, not surprisingly, are a lot less likely to be validated.

The Free tags displayed the pros and cons of folksonomies well, in that almost all the errors, all the obscene tags, alongside the majority of terms from the remaining subjective term-categories not represented in descriptors or validated tags, was found within the Free tags. As reported by Bischoff et al. (2008) subjective opinions/qualities were used quite often as queries in folksonomies, but tended to be neglected as a tag-type. This could make an argument for some utilization of the Free tags as they could provide this kind of access point to an image collection.

4.3 RQ2 findings and discussion

RQ2: To what extent do taggers and indexers use the same terms to describe the same image?

To do the automatic matching as described in the previous chapter (3.3.3), each tag had to be coupled in excel with each corresponding descriptor. Table 9 displays the total number of tag-descriptor pairs across each of the three Tag-datasets included in the analysis.

Table 9
Tag-descriptors pairs across the three datasets

	Descriptors (n=7306), Images (n=2079)		
	Free tags (n=15525), Images (n=2079)	2VTags (n=4743), Images (n=1864)	3VTags (n=2516), Images (n=1501)
Tag-descriptor pairs	55634	16886	9061

Table 10
Exact and fuzzy frequency and overlap between descriptors and: Free tags, 2Vtags and 3Vtags

		Exact matches	Fuzzy matches
Free tags	Frequency (Overlap %)	1026 (4,7%)	2591 (12,8%)
	Mean (Standard Deviation)	0,5 (0,74)	1,26 (1,24)
	Max	4	7
2Vtags	Frequency (Overlap %)	568 (4,9%)	1392 (13,1%)
	Mean (Standard Deviation)	0,3 (0,56)	0,75 (0,82)
	Max	3	5
3Vtags	Frequency (Overlap %)	368 (3,89%)	917 (10,3%)
	Mean (Standard Deviation)	0,25 (0,49)	0,61 (0,69)
	Max	3	4

We find rather low values across the board, which in terms of overlap most closely resembles the investigation between tags aggregated from different folksonomies and LCSH done by Thomas, Caudle, & Schmitz (2009), in which they exact overlap of 6%, and the study by Lykke, Hoj, Madsen, Golub, & Tudhope (2012) that found this type of overlap to between 3-4,4%.

We can conclude that fuzzy overlap, not surprisingly, is a more likely to occur than exact. An interesting observation is that fuzzy matches occur more than twice as often as the character-for-character exact match, showing that taggers and indexers choose different tenses often; or, rather, that taggers tend to use singular (as the descriptors always will be in plural). When thinking about the simple cognitive process behind tagging (2.4.5), one can imagine how a player would add singular tags – e.g. if they see just one swimsuit add the tag 'swimsuit', whereas the descriptors uses the plural 'swimsuit(s)', even though the photograph only depicts one swimsuit. In figure 36 below, a complete example of matches is shown – 1 exact and 3 fuzzy. We see how only the term Fiskeri (fishing) matches and the three other terms (harbor, cap and fishing boat) all use different forms.

Figure 36 – Examples of three fuzzy and one exact match

The overlap is illustrated with Venn diagrams in figure 37 below:

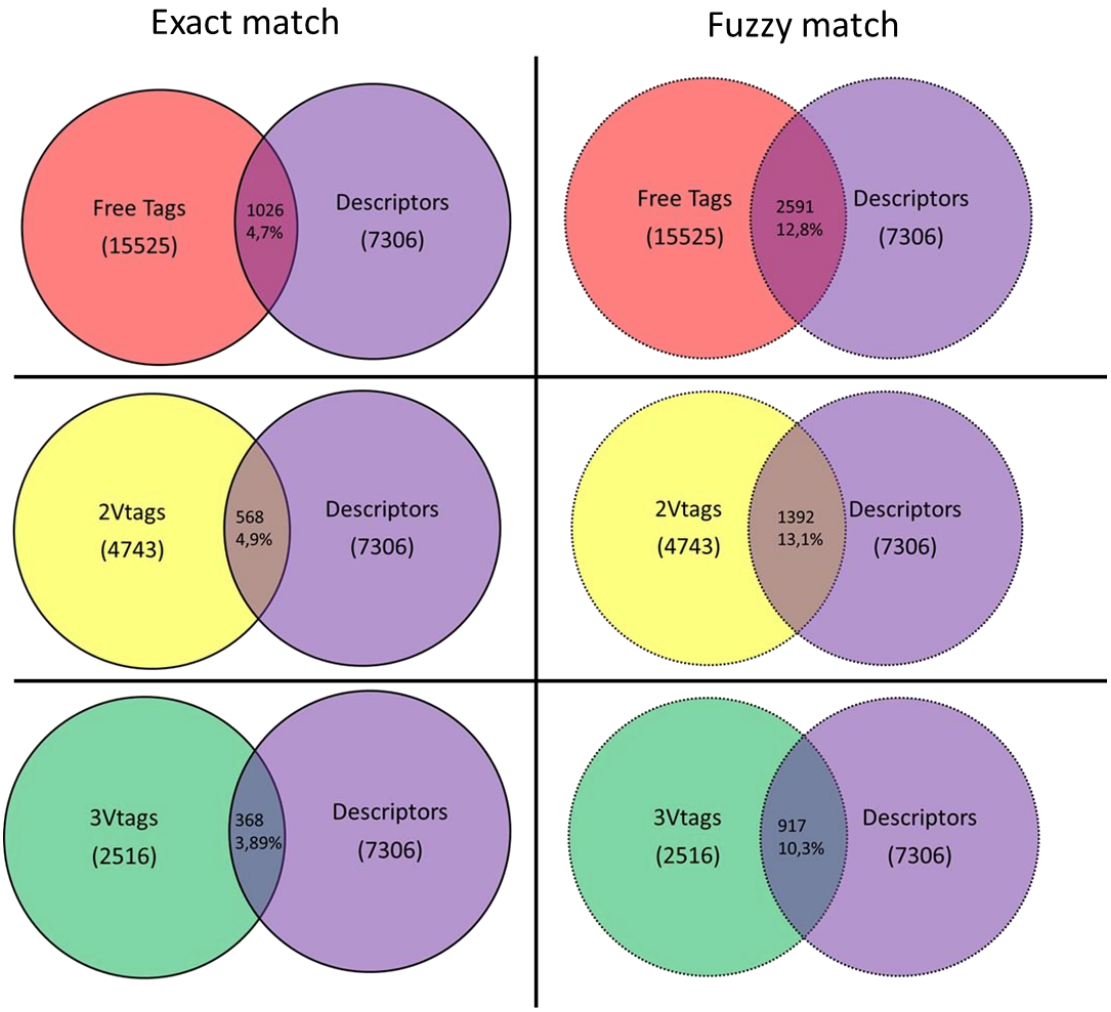

Figure 37 - Venn diagrams with exact and fuzzy overlap

We can conclude that the degree of overlap doesn't seem to change in any consistent way with a stricter validation threshold – there is a slight drop in overlap between 2Vtags and 3Vtags and the 2Vtags are slightly more similar to the descriptors than the Free tags and the 3Vtags.

Overall the Venn diagrams show a low similarity among tags and descriptors. If we consider the much lower numbers of 2Vtags and 3Vtags, however, the overlap shows that the validation process equals fewer tags – but that those are more likely to have a syntactic overlap with the descriptors as seen in table 11:

Table 11
Percentage of tags with fuzzy match with descriptors

	Free tags (n=15525)	2Vtags (n=746)	3Vtags (n=2516)
Frequency of fuzzy matches	2591 (16,68%)	1392 (29,34%)	917 (36,44%)

If we view these frequencies of fuzzy matches between terms as something similar to the discussion on inter-indexing consistency (chapter 2.3.5 & 2.6.2), seeing the total sum of players as one indexer and the indexers at KB as another, we can compare the inter-indexer consistency reported for a similar collection with two indexers using the same ontology (Hughes & Rafferty, 2011) was 27% or almost the same as the 2Vtags. The numbers aren't directly comparable as the method used in that study was different, but the frequency of fuzzy matches compared to the total number of tags, especially for the 2Vtags and 3Vtags, shows that the game-generated tags are reasonably good at agreeing with one of the descriptors.

4.4 RQ3 findings and discussion

RQ3: To what extent do taggers and indexers use thesaurus-related terms to describe the same image?

The sample of 326 out of 2079 images was chosen and found to contain:

Table 12

Number of terms in category analysis sample

Free tags	2Vtags	3Vtags	Descriptors
2480	746	380	1112

The relation-types were then coded according the method described in section 3.3.3 by comparing each tag to the corresponding set of descriptors and identifying thesaurus relations.

These findings are a natural extension of the results found in RQ2; first, I investigated whether tagger and indexer use the exact same words, followed up by variations of the same words, and then I widened the searchlight to look for overlap in meaning.

Below are the findings in three tables; one or for Free tags and descriptors (table 13), one for 2Vtags and descriptors (table 14) and one for 3Vtags and descriptors (table 15).

The first column shows the number of times each relation was established. The percentage for the total semantic overlap is the Jaccard coeffiecient calculated on the basis of the numbers in table 12 and the total number of tags that had a thesaurus relation with a descriptor (the total semantic overlap)

The second column shows how large a percentage of the total semantic overlap each relation type constitutes.

The third column shows the mean and standard deviation on object level i.e. how often a certain Relation type is established between tags and descriptors on average for each image.

Table 13
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)
Tag-literal	52	5,73 %	0,16 (0,47)
Literal-descriptor	13	1,43 %	0,04 (0,25)
Associative	250	27,56 %	0,77 (1,34)
Total semantic overlap	907 (33,78%)	100 %	0,36 (0,48)

Table 14
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)
Tag-literal	20	5,32%	0,07 (0,28)
Literal-descriptor	2	0,53%	0,01 (0,08)
Associative	74	19,68%	0,25 (0,61)
Total semantic overlap	376 (25,37%)	100%	0,49 (0,50)

Table 15
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)
Tag-literal	8	3,74 %	0,03 (0,18)
Literal-descriptor	2	0,93 %	0,01 (0,09)
Associative	40	18,69 %	0,17 (0,4)
Total semantic overlap	214 (16,74%)	100 %	0,56 (0,49)

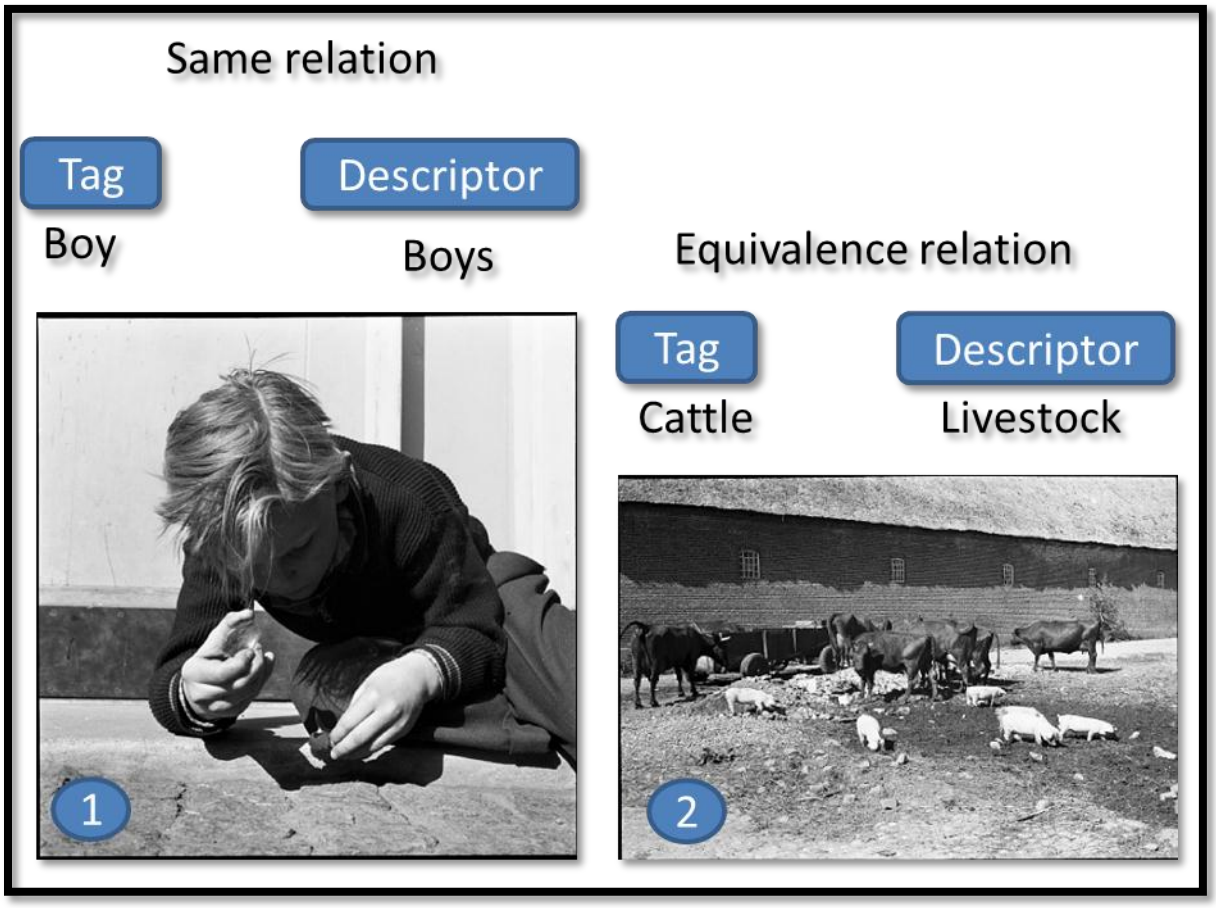

Figure 38 – Same and equivalence relations

1: Same-relation established between the tag and descriptors sharing the same word-stem. This relation was established 365, 205, 135 times and is the most frequently occurring of all thesaurus relations across all three types of tags. The main thing to notice is agreement between the algorithm applied to the whole set (RQ2) and the human analysis done on a smaller sample. The Jaccard coefficient for the two methods is listed in table 16:

Table 16
Same relation (syntactic fuzzy match %)

	Algorithm on whole dataset (n=2079)	Human validation on sample (n=326)
Free Tags	12,8 %	11,3 %
2Vtags	13,1 %	12,4 %
3Vtags	10,3 %	9,7 %

The slightly higher match for the algorithm can be explained by the false positives (see chapter 3.3.2.3).

2: The 'strongest' thesaurus-relation, in that tagger and indexer wanted to express the same subject but with different terms, and one of the ones which almost never occurs in the data. In the rare cases when players and indexers want to express the exact same subject, they choose the same

exact term, rather than synonyms. The low numbers are very similar to the findings for synonyms found in other studies between tags and descriptors (Lykke, Høj, Madsen, Golub, & Tudhope, 2012).

Figure 39 – Broader/narrower relation

1: Tree (træ) is a broader term of the more specific Christmas tree (juletræ).

Broader relationship was the most frequently established kind of hierarchical thesaurus-relation. It was observed 74 times for the Free tags, 33 times for 2Vtags and 17 times for 3Vtags.

2: Evening dress (aftenkjole) is narrower term of the less specific dress (kjole).

Narrower was more rare and was established 54 times for Free tags, 11 times for 2Vtags and 7 times for 3Vtags.

Players are a more likely to describe things at a general level, compared to indexers who tend to be a more specific, in these cases qualifying what *kind* of tree or dress is shown.

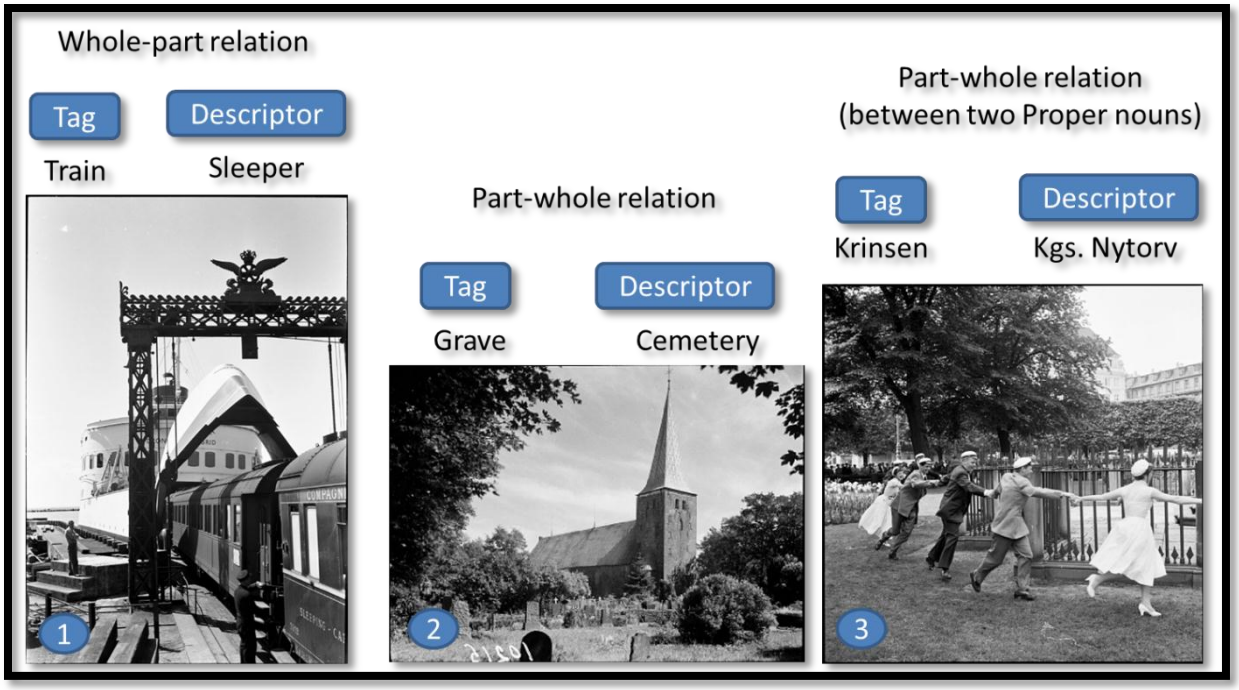

Figure 40 – Whole-part/part-whole relation

1: The tag refers to the train (tog) in the image and the descriptor to the special kind of train cart seen in the image, called a sleeper (sovevogn) – creating a whole-part relationship.

53 of the Free tags, 6 of the 2Vtags and 3 of the 3Vtags had a Whole-part relation to the descriptors.

2: The tag refers to a grave (gravsted) and the descriptor to a cemetery (kirkegård), making it a part-whole relationship.

3: The descriptor refers to the square (Kgs. Nytorv) where the picture is taken and the tag refers specifically to the green area (Krinsen) in the middle of the square, providing an example of the part-whole relationship between two Proper nouns.

9 of the Free tags, 6 of the 2Vtags and 2 of the 3Vtags had a Part-whole relation to the descriptors.

Aside from the whole-part relations between Free tags and descriptors – which was 52 or 5,84% of the total semantic overlap between those datasets, these relations were very rare.

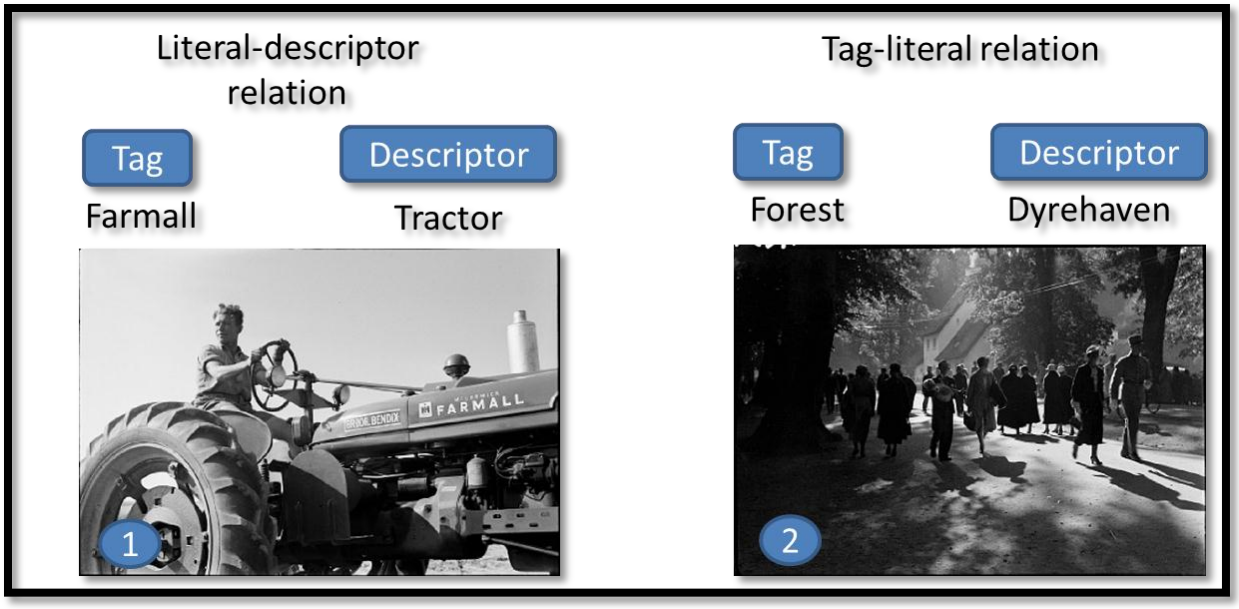

Figure 41 – Literal-descriptor/Tag-literal relation

1: An example of the 'From image' term-category, where the tag 'Farmall' is an instance of the descriptor 'Tractor' (traktor).

This relation was very rare and only occurred 13 times for the Free tags. Other examples were cities, where the descriptor merely read 'town' and the player had recognized which town and assigned the proper noun. Only 2 of these relations were noted for 2Vtags and 3Vtags.

2: The descriptor 'Dyrehaven' is the name of a forest area north of Copenhagen and the tag 'Forest (skov) forms a Tag-literal relation to the descriptor.

This relation was established more frequently, with 52 for the Free tags, 20 for the 2Vtags and 8 for the 3Vtags. Considering the findings in RQ1 relating to the more widespread use of 'Location' in descriptors, this is not surprising. Tag-literal is typically a place, named in the descriptors and then described at a general level in the tags e.g. forest, city, street or park.

As "Broader + Whole-part + Tag-literal" relations indicates a tag at a more general level than the descriptor and "Narrower + Part-whole + Literal-descriptor" relations signifies a tag at a more specific level, the hierarchical thesaurus relations can be grouped together like seen in table 17:

Table 17
Hierachical relations

	Free tags	2Vtags	3Vtags
Broader + Whole-part + Tag-literal	179	66	28
Narrower + Part-whole + Literal-descriptor	76	19	11

To show how tags – when there is a thesaurus match - describe more general concepts than the descriptors.

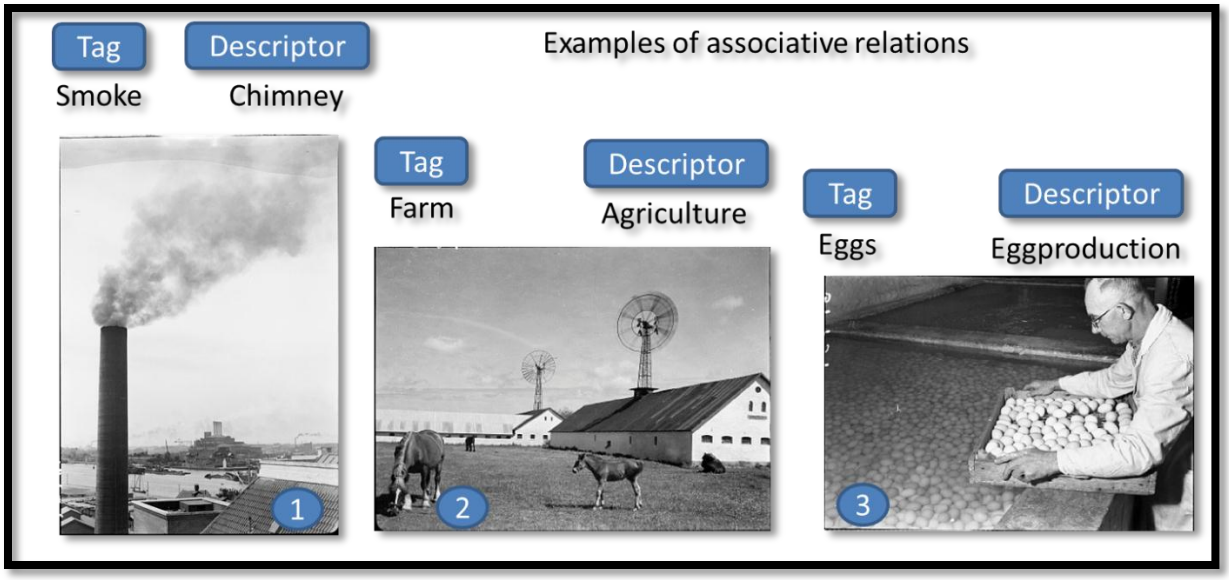

Figure 42 – Associative relations

1: Example of an associative relation, where the definition of a 'chimney' (skorsten) is that is channels 'smoke' (røg).

2: Example of an associative relation where the tag 'farm' (gård) belongs to the disclipline 'agriculture'.

3: Example of an associative relation where the action 'eggproduction' (æggeproduktion) results in the product 'eggs' (æg).

As explained in the methodology (chapter 3.3.3.1), the associative relations are rather vaguely defined as the three examples from figure 42 shows with their variety. Even with the strict 'one degree of separation' 250 (Free tags), 74 (2Vtags) and 40 (3Vtags) had this relation to the descriptors and was the second most frequently occurring thesaurus relation.

Overall these different thesaurus-relations between tags and descriptors show an overlap of meaning, which goes beyond simple term overlap. We can conclude that the overlap – when looking at meaning instead of just terms – rose substantially for both Free tags, 2Vtags and 3Vtags.

For the Free tags there were 365 cases of syntactic overlap and 907 semantic in the sample – or almost three times as many relations. For the 2Vtags there were 205 cases of syntactic overlap and 376 semantic – almost twice as many. For the 3Vtags there were 132 cases of syntactic overlap and 214 semantic, signifying a more than 60% increase in the number of relations.

In chapter 3.3.3.1 the overarching thesaurus relations were listed as:

- Equivalence
- Hierarchical
- Associative

The hierarchical relations were analyzed with granularity by differentiating between 'type of' (Broader/narrower), 'part of' (Whole-part/part-whole) and 'instance of' (Literal-descriptor/descriptor-literal). Grouping them back with each other, while also grouping 'Same' with 'Equivalence' lets us illustrate how the overlap is broken down among the different types of tags.

In figure 43 we see the 907 thesaurus relations between the Free tags and the descriptors.

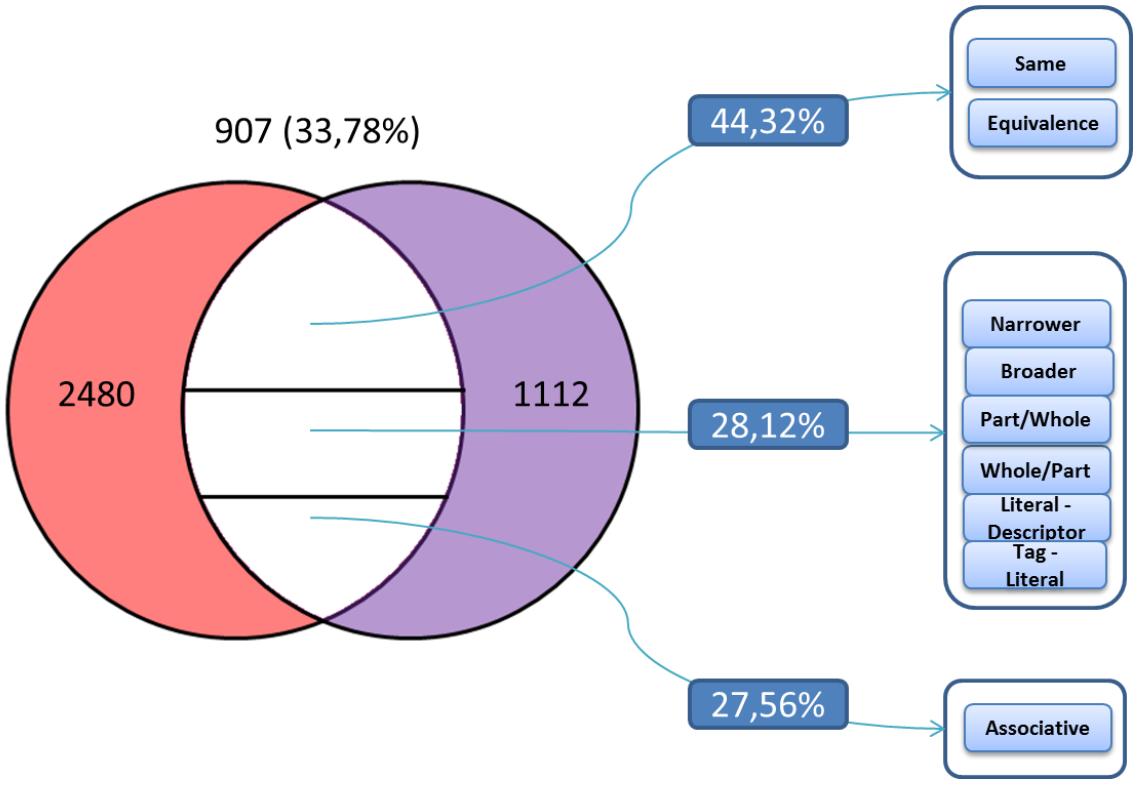

Figure 43 – Thesaurus relations between Free tags and descriptors

In figure 44 we see the 376 thesaurus relations between the 2Vtags and the descriptors:

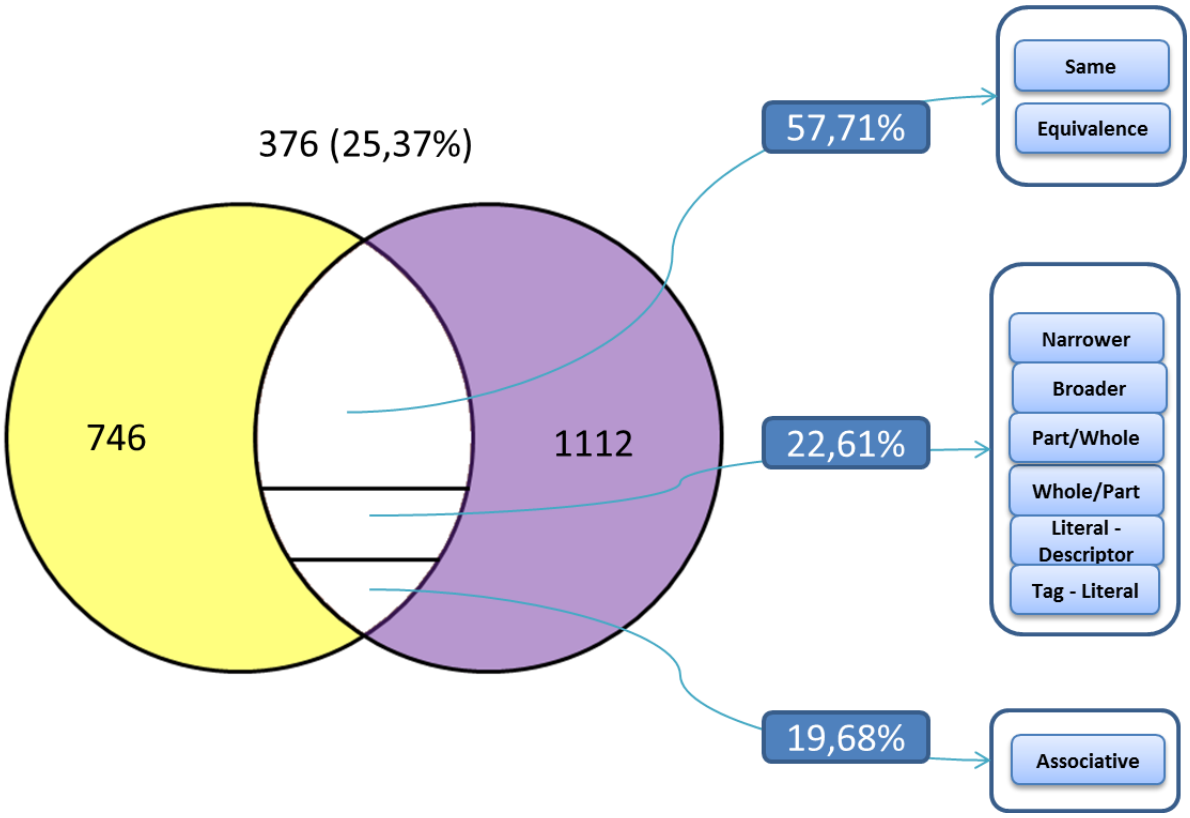

Figure 44 - Thesaurus relations between 2Vtags and descriptors

In figure 45 we see the 214 thesaurus relations between the 3Vtags and the descriptors:

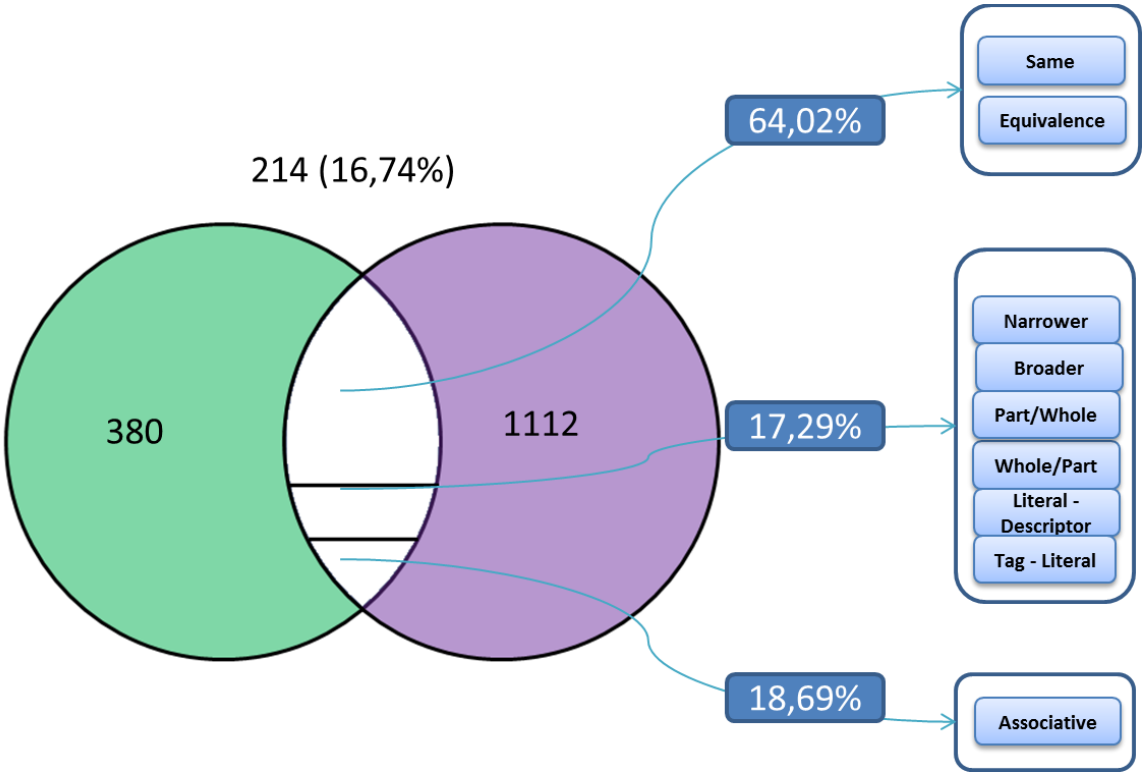

Figure 45- Thesaurus relations between 3Vtags and descriptors

A higher proportion of the Free tags belonged to the hierarchical and associative types of relation, than for both 2Vtags and 3Vtags, that mainly belonged to the equivalent (and Same) type.

When looking at the number of established relations compared to the number of tags in each group, the percentage of tags with some sort of relationship to the descriptors increases with the validation threshold:

Table 18
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%)

This shows a strong relationship between the underlying meaning expressed by the output of the GeF game and the meaning expressed by indexers. More than half the validated tags have a thesaurus relation to a descriptor at object level and the same is true for more than a third of the Free tags.

Hierarchical and associative relationships respectively made up roughly the same portion of the semantic overlap for each set of tags, but the Free tags had a lower proportion of 'Same and equivalence' and therefore a higher amount of both hierarchical and associative relations.

Chapter 5 Conclusion

5.1 Conclusion

RQ1: To what extent do tags (across all three validation thresholds) and descriptors fall within the same term-categories?

The Free tags displayed the pros and cons of folksonomies well, in that almost all the errors and all obscene tags was from that dataset - alongside the majority of terms not represented in descriptors or validated tags (From image, Time, Modern and Subjective/narrative).

The professionally assigned descriptors were primarily 'Artifact/objects' and 'Proper noun' with a smaller portion from the 'Actions/events' term-category. Only a very few descriptors were subjective/narrative.

The 2Vtags and 3Vtags were mainly 'Artifacts/objects'-tags and the overweight of 2Vtags and 3Vtags in this term-category shows that this type of term, which represents what is *in* the picture, rather than what it is *about* are most likely to be validated in a Human Computation output agreement game like GeF.

RQ2: To what extent do taggers and indexers use the same terms to describe the same image?

Via automatic means it was determined that the exact (character-for-character) overlap i.e. the number of common terms compared to the entire pool of tags and descriptors was slightly less than 5% for all three types of tags. By extending the analysis to include fuzzy (word-stem) matching, the overlap more than doubled.

Looking just at how many tags that had a fuzzy match with a descriptor, the percentages were 16,68% for Free tags, 29,34% for 2Vtags and 36,44% for 3Vtags.

RQ3: To what extent do taggers and indexers use thesaurus-related terms to describe the same image?

This expanded on RQ2, by investigating not only the syntactic overlap of the terms, but the overlap in meaning (semantic), as well.

It was shown that that the overlap between all three types of tags and the descriptors rose substantially when taking this more inclusive understanding of overlap.

The same was true for the percentages of tags matching a descriptor, which rose to 36,57% for Free tags, 50,40% for 2Vtags and 56,31% for 3Vtags.

The 'Same' (term match) was the most frequently observed relation in the sample, but both hierarchical and associative relations were consistently represented for all three kinds of tags. When hierarchical relations occur, tags tend to be at a more general level than descriptors.

5.2 Implications for future research

The data analyzed in this thesis could lend itself well to further studies involving users and relevance assessment across validation thresholds and term-categories. It was shown that the Free tags hold the majority of 'Subjective/narrative' terms. Investigating whether these terms increase retrieval or facilitate better browsing could be interesting paths to explore.

In addition, it could be interesting to see how the different datasets compare to each other in general relevance tests. The creators of the ESP-game did an evaluation of the output of their game (chapter 2.7.4) and found that their validated tags were deemed relevant. Testing on real users to see if game-generated tags actually are as good (or better) entry-points to a collection as the expensive, professional descriptors could inform future indexing policy at institutions willing to invest time in developing crowdsourcing games.

The Sven Türck collection is just one out of many housed at KB. Running similar games with different types of images, such as maps, newspaper-illustrations or artwork might yield different results.

The method of comparing thesaurus relations has to my knowledge not been applied to image subject terms before and it turned out to be a useful way to determine semantic overlap between two types. It could potentially be applied to other cases where image collections are described by different means – for example by comparing images that have been tagged in Flickr: The Commons with metadata from the libraries and museums that have uploaded the images.

5.3 Recommendations

The findings indicate that a game like GeF could potentially supplement or perhaps even replace part of the in-house indexing done at institutions with image collections in need of subject metadata. While the overlap measured between the validated tags and descriptors was low due to the low number of validated tags, the percentage of validated tags that had some sort of relation to a corresponding descriptor was over 50%.

As almost 90% of the 3Vtags belong to the 'Artifacts/objects' term category these are also the types of terms that will be common between tags and descriptors. 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 kinds of metadata would be to let the indexers add 'Proper nouns' (mainly locations) and let the players add information about 'Artifacts/objects' as the game lends itself well to those sort of descriptions.

In the case of the Sven Türck collection the 3Vtags alone would not add much to the existing metadata, as there is only 1,66 tags on average for each picture, with a third of these already in the descriptors. The 2Vtags have almost the same characteristics; they are from the same term-category and have the same thesaurus relations to the descriptors, but there are almost twice as many of them.

An even more radical approach would be to simply use all Free tags generated. 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. There are two ways to deal with this:

The problem with errors/abuses can either be fixed before or after the tagging occurs. Pre-tag cleaning would entail a mechanism of auto-correction based on either a dictionary or some existing controlled vocabulary – that only allows controlled terms to be entered, which might rob the final outcome of some of the more creative tags.

Post-tag cleaning could take the form of screening on vocabulary level rather than object level. It was shown in this study that a high percentage of the tags actually shared meaning with the descriptors. If it is assumed that the descriptors are 'correct' it can be inferred that the same must also be true for the tags. If this is accepted, the indexer needs only to look at unique tags to weed out the rare obscene tags and correct/delete spelling mistakes, as no tag-image analysis is needed. This would be a feasible (speedy) way to enrich a collection with the diverse perspectives from the folksonomies.

References

Al-Khalifa, H., & Davis, H. (2006). Folksonomies versus automatic keyword extraction: an empirical study. *IADIS International Journal on Computer Science and Information Systems, 1*(2), 132-143.

Al-Khalifa, H., & Davis, H. (2007). Towards Better Understanding of Folksonomic Patterns. *Proceedings of the 18th Conference on Hypertext and Hypermedia* (p. 163). Manchester: ACM.

Andersen, T. J. (2010, November 23). Hvad forestiller billedet? Nyt FB-spil fra Det Kongelige Bibliotek. (A. H. Nissen, Interviewer) P1.

Anderson, C. (2004). The Long Tail. *Wired magazine, 12*(10).

Bainbridge, D., Twidale, M., & Nichols, D. (2011). That's' {\'e}'not'{\th}''?'or'□': a user-driven context-aware approach to erroneous metadata in digital libraries. *JCDL '11* (pp. 39-48). Ottawa, Canada: ACM.

Bates, M. (1998). Indexing and access for digital libraries and the Internet. *Journal of the American Society for Information Science, 49*, 1185-1205.

Beaudoin, J. (2007). Flickr image tagging: patterns made visible. *Bulletin of the ASIST, 34*(1), 26-29.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? *CIKM 2008 Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 203-212). Napa Valley, California: ACM.

Brooks, T., Carli, S., Dallmeier-Tiessen, S., Mele, S., & Weiler, H. (2011). Authormagic in INSPIRE - author disambiguation in scholarly communication. *Proceedings of the ACM WebSci '11* (pp. 1-2). Koblenz, Germany: ACM.

Brown, P., Hidderley, R., Griffin, H., & Rollason, S. (1996). The democratic indexing of images. *New Review of Hypermedia and Multimedia: Applications and Research, 2*(1), 107-120.

Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. *Proc. ISWC 2008*, *5318 of LNCS*, pp. 615-631. Karlsruhe.

Chen, H.-l. (2001). An analysis of image queries in the field of art history. *Journal for the American Society for Information Science and Technology, 52*(3), 260-273.

Christensen-Dalsgaard, B. (2010). Crowdsurfing - hvem, hvad og hvorfor? *Revy, 33*(6), 20-21.

Conradi, E. (2011). to_be_classified. *Journal of Information Architecture, 2*(2), 5-24.

Cutter, C. A. (1876). *Rules for a Printed Dictionary Catalog.* Washington: U.S Bureau of Education.

David, C., Giroux, L., Bertrand-Gastaldy, S., Lanteigne, D., & Bertrand, A. (1995). Indexing as problem solving: a cognitive approach to consistency. *Proceedings of the annual meeting-american society for information science. 32*, pp. 49-55. Learned Information (EUROPE).

Dawson, R. (2010, May 29). *Crowdsourcing results.* Retrieved June 20, 2012, from http://crowdsourcingresults.com/competition-platforms/crowdsourcing-landscape-discussion/

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

Enser, P. (2007). Visual Image Retrieval. In *Annual Review of Information Science and Technology (ARIST) Vol 42* (Vol. 42, pp. 3-42). New York: American Association of Information Science and Technology.

Enser, P., & McGregor, C. (1992). *Analysis of Visual Information Retrieval Queries.* London: British Library.

Enser, P., Sandom, C., & Lewis, P. H. (2005). Surveying the reality of semantic image retrieval. *Visual 2005: 8th International Conference on Visual Information Systems*, (pp. 177-188). Amsterdam.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science, 45*(8), 572-576.

Giles, J. (2005). Special Report Internet encyclopaedias go head to head. *Nature, 438*, 900-901.

Gligorov, R., Baltussen, L., Ossenbruggen, J., Aroyo, L., Brinkerik, M., Oomen, J., et al. (2010). Towards Integration of End User Tags with Professional Annotations. *Web Science Conf. 2010*, (pp. 1-6). Raleigh, NC.

Golder, S., & Huberman, B. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter, 12*(1), 58-72.

Guy, M., & Tonkin, E. (2006). Folksonomies Tidying up tags? *D-Lib Magazine, 12*(1).

Hassan-Montero, Y., & Herrero-Solana, V. (2006). ImprovING TAG-CLOUDS AS VISUAL INFORMATION RETRIEVAL INTERFACES. *International Conference on Multidisciplinary Information Sciences and Technologies.* Mérida, Spain: InSciT.

Heckner, M., Mühlbacher, S., & Wolff, C. (2008). Tagging tagging. Analysing user keywords in scientific bibliography management systems. *Journal of Digital Information, 9*(2).

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine, 16*(3/4).

Howe, J. (2006). The Rise of Crowdsourcing. *Wired magazine, 14*(14), 1-5.

Hughes, A. V., & Rafferty, P. (2011). Inter-indexer consistency in graphic materials indexing at the National Library of Wales. *Journal of Documentation, 67*(1), 9-32.

Ingwersen, P. (2002). Cognitive perspectives of document representation. *CoLIS 4: 4th International Conference on Conceptions of Library and Information Science* (pp. 285-300). Seattle: Libraries Unlimited.

Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., & Stumme, S. (2008). Discovering shared conceptualizations in folksonomies. *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(1), 38-53.

Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management, 34*(2-3), 161-174.

Jörgensen, C. (2007). Image Access, the Semantic Gap, and Social Tagging as a Paradigm Shift. *18th Annual ASIS SIG/CR Classification Research Workshop.* Milwaukee, Wisconsin.

Kipp, M. (2005). 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.

Krause, M. (1988). Intellectual problems of indexing picture collections. *Audiovisual librarian*, 73-81.

Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice* (3rd ed.). London: Facet Publishing.

Law, E., & von Ahn, L. (2009). Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games. *CHI '09 Proceedings of the 27th international conference on Human factors in computing systems* (pp. 1197-1206). Boston: ACM.

Layne, S. (1994). Some Issues in the Indexing of Images. *Journal of the American Society for Information Science, 45*(8), 583-588.

Lee, D., & Schleyer, T. (2010). A comparison of meSH terms and CiteULike social tags as metadata for the same items. *IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 445-448). New York: ACM.

Leonard, L. (1977). *Inter-indexer consistency studies, 1954-1975: A review of the literature and summary of study results.* Urbana-Champaign: University of Illinois. Graduate School of Library Science.

Lu, C., Park, J.-r., & 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.

Lykke, M., Høj, A., Madsen, L., Golub, K., & Tudhope, D. (2012). Tagging behaviour with support from controlled vocabulary. *Proceedings 2nd biennial Conference (Facets of KnowledgProceedings 2nd biennial Conference of the British Chapter of the International Society for Knowledge Organization* (pp. 41-50). London: Emerald Group Publishing Limited.

Maness, J. (2006). Library 2.0 Theory: Web 2.0 and Its Implications for Libraries. *Webology, 3*(2).

Markey, K. (1984). Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research, 6*, 155-177.

Marlow, C., Naaman, M., boyd, d., & Davis, N. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. *Proceedings of the 17th Conference on Hypertext and Hypermedia*, (pp. 31-40). Odense, Denmark.

Marshall, C., & Rossman, G. (2006). *Designing Qualitative Research.* Thousand Oaks: Sage Publication.

Mathes, A. (2004). Retrieved 05 18, 2012, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Matthews, J. (2000). The Value of Information in Library Catalogs. *Information Outlook, 4*(7), 18-24.

Matusiak, K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services: International digital library perspectives, 22*(4), 283-298.

Medelyan, O., & Witten, I. (2006). Measuring inter-indexer consistency using a thesaurus. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 274-275). Chapel Hill, NC: ACM.

Ménard, E. (2007). Indexing and retrieving images in a multilingual world. *Knowledge Organization, 34*(2), 91-100.

Moltved, N. K. (2011, January 25). Retrieved February 25, 2012, from Arkivformidling: http://arkivformidling.wordpress.com/2011/01/25/fa-tagget-dine-billeder-i-et-ruf/

Motter, A., Moura, A., Lai, Y.-C., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E, 65*, 065102.

Moulaison, H. (2008). Social tagging in the web 2.0 environment: author vs. user tagging. *Journal of Library Metadata, 8*(2), 101-111.

Müller, R., Thoring, K., & Oostinga, R. (2010). Crowdsourcing with Semantic Differentials: A Game to Investigate the Meaning of Form. *AMCIS 2010 Proceedings. Paper 342.* Americas Conference on Information Systems (AMCIS).

Nagarajan, M., Verma, K., Sheth, A., & Miller, J. L. (2006). Semantic Interoperability of Web Services – Challenges and Experiences. *IEEE International Conference on Web Services (ICWS'06)* (pp. 373 - 382). Chicago: IEEE.

Oomen, J., & Aroyo, L. (2011). Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges. *5th International Conference on Communities and Technologies* . Brisbane, Australia.

Ørnager, S. (1999, Juli). Billeder og Ord - Analyse, beskrivelse og -søgning af pressefoto. Handelshøjskolen København.

Panofsky, E. (1970). *Meaning in the visual arts.* London: Penguin.

Peters, I. (2009). *Folksonomies : Indexing and Retrieval in Web 2.0.* Berlin: De Gruyter.

Peters, I., & Stock, W. (2010). "Power tags" in information retrieval. *Library Hi Tech, 28*(1), 81-93.

Peters, I., Schumann, L., Terliesner, J., & Stock, W. (2011). Retrieval effectiveness of tagging systems. *Proceedings of the American Society for Information Science and Technology. 48*, pp. 1-4. American Society for Information Science and Technology.

Quintarelli, E. (2005). Folksonomies : power to the people. *International society for knowledge organization Italy-University of Milano Bicocca - ISKO Italy-UniMiB meeting.* Milan.

Rasmussen, E. (1997). Indexing images. In *Annual Review of Information Science and Technology (ARIST) Vol 32* (pp. 169-196). New York: American Association of Information Science and Technology.

Rorissa, A. (2010). A comparative study of Flickr tags and index terms in a general image collection. *Journal of the American Society for Information Science and Technology, 61*(11), 2230-2242.

Schmitz, C., Baldassarri, A., Servedio, V., Loreto, V., Hotho, A., Grahl, M., et al. (2007). Network properties of folksonomies. *AI Communications - Network Analysis in Natural Sciences and Engineering, 20*(4), 245-262.

Shatford, S. (1986). Analyzing the subject of a picture: a theoretical approach. *Cataloging & Classification Quarterly, 6*(3), 39-62.

Shirky, C. (2010). *Cognitive Surplus: Creativity and Generosity in a Connected Age .* London: Penguin Group.

Sinha, R. (2005, September 27). *A cognitive analysis of tagging*. Retrieved May 12, 2012, from http://rashmisinha.com: http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/

Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349-1380.

Springer, M. (2008). *For the Common Good: The Library of Congress Flickr Pilot Project.* Washington: Library of Congress (LOC).

Steele, T. (2009). The new cooperative cataloging. *Library Hi Tech, 27*(1), 68-77.

Stock, W. (2007). Folksonomies and science communicationA mash-up of professional science databases and Web 2.0 services. *Information Services and Use, 27*(3), 97-103.

Svenonius, E. (2000). *The intellectual foundation of information organization.* Cambridge: MIT.

Taylor, A., & Joudrey, D. N. (2009). *The organization of information* (3rd ed.). Westport: Libraries Unlimited.

Tennant, R. (1998). Digital Libraries: 21st-Century Cataloging. *Library Journal, 123*(7), 1998.

Thomas, D. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation, 27*(2), 237-246.

Thomas, M., Caudle, D., & Schmitz, C. (2009). To tag or not to tag? *Library Hi Tech, 27*(3), 411-434.

Trant, J. (2006). Exploring the potential for social tagging and folksonomy in art museums: proof of concept. *New review of Hypermedia and Multimedia, 12*(1), 83-105.

Vander Wal, T. (2005). *Explaining and Showing Broad and Narrow Folksonomies*. Retrieved February 3, 2012, from Personal InfoCloud: http://personalinfocloud.com/2005/02/explaining_and_.html

Veres, C. (2006). Concept modeling by the masses: folksonomy structure and interoperability. In D. Embley, A. Olive, & R. Sudha (Eds.), *Conceptual Modeling - ER 2006* (pp. 325-338). Springer Berlin / Heidelberg.

von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine, 39*(6), 96-98.

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proc. SIGCHI Conf. on Human Factors in Computing Systems* (pp. 319-326). Vienna: ACM.

Voorbij, H. (1998). 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.

Wagner, C. (2006). Breaking the Knowledge Acquisition Bottleneck Through Conversational Knowledge Management. *Information Resources Management Journal, 19*(1), 70-83.

Weber, R. (1990). *Basic content analysis.* categories is a form of content analysis, typically divided into the creation of a coding scheme and a definition of the recording units, then an assessment of the accuracy of the coding on a smaller sample, a revision and finally coding the entire text : Sage Publications.

Wetterström, M. (2008). 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.

Yi, K., & Chan, L. M. (2009). Linking folksonomy to Library of Congress subject headings: an exploratory study. *Journal of Documentation, 65*(6), 872-900.

Yoon, J., & O'Connor, B. (2010). Engineering an image-browsing environment: re-purposing existing denotative descriptors. *Journal of Documentation, 66*(5), 750-774.

Yuen, M., Chen, L., & King, I. (2009). A survey of human computation systems. *International Conference on Computational Science and Engineering* (pp. 723-728). Vancouver, Canada: IEEE.

Appendices

Appendix A: Sample original MODS metadata

<record><header><identifier>oai:kb.dk:images:billed:2010:okt:billeder:object76114</identifier>
<datestamp>1327487976000</datestamp><setSpec>oai:kb.dk:images:billed:2010:okt:billeder</setSpec></head
er><metadata><md:mods xmlns:md="http://www.loc.gov/mods/v3"
xmlns:java="http://xml.apache.org/xalan/java" xmlns:t="http://www.tei-c.org/ns/1.0"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-3.xsd">
<!--Generated by COP mods_generator.xsl billeder $Id: mods_generator.xsl,v 1.29 2011-03-29 08:04:50 slu Exp
$
--><md:recordInfo><md:languageOfCataloging><md:languageTerm authority="rfc4646"/>
</md:languageOfCataloging>
<md:recordIdentifier>/images/billed/2010/okt/billeder/object76114</md:recordIdentifier>
<md:recordCreationDate encoding="w3cdtf">2007-03-26</md:recordCreationDate>
<md:recordChangeDate encoding="w3cdtf">2011-01-19</md:recordChangeDate></md:recordInfo>
<md:titleInfo xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="">
<md:title>Gummibåd</md:title></md:titleInfo>
<md:name type="personal" xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="">
<md:namePart>Türck, Sven (1897-1954) fotograf</md:namePart>
<md:role><md:roleTerm type="text">creator</md:roleTerm></md:role></md:name>
<md:name type="cumulus" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang=""><md:namePart>cumulus:cumulus</md:namePart>
<md:role><md:roleTerm type="text">last-modified-by</md:roleTerm></md:role></md:name>
<md:relatedItem type="event"/><md:extension><!-- node=668-->
<h:div xmlns:h="http://www.w3.org/1999/xhtml" xlink:href="#668">
<h:a href="http://www.kb.dk/editions/any/2009/jul/editions/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Home</h:a>
<h:a href="http://www.kb.dk/editions/any/2009/jul/editions/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Hjem</h:a>
/ <h:a href="http://www.kb.dk/subject2108/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Billeder</h:a>
<h:a href="http://www.kb.dk/subject2108/da/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">Billeder</h:a> / <h:a href="http://www.kb.dk/subject2109/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Samlinger</h:a><h:a
href="http://www.kb.dk/subject2109/da/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">Samlinger</h:a> / <h:a href="http://www.kb.dk/subject2112/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Fotografarkiver</h:a><h:a
href="http://www.kb.dk/subject2112/da/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">Fotografarkiver</h:a> / <h:a href="http://www.kb.dk/subject668/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Türck, Sven</h:a><h:a
href="http://www.kb.dk/subject668/da/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">Türck, Sven</h:a></h:div></md:extension><md:extension><!--
node=673--><h:div xmlns:h="http://www.w3.org/1999/xhtml" xlink:href="#673"><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Home</h:a><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Hjem</h:a> / <h:a

href="http://www.kb.dk/subject2108/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Billeder</h:a><h:a href="http://www.kb.dk/subject2108/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Billeder</h:a> / <h:a
href="http://www.kb.dk/subject2109/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Samlinger</h:a><h:a href="http://www.kb.dk/subject2109/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Samlinger</h:a> / <h:a
href="http://www.kb.dk/subject2112/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Fotografarkiver</h:a><h:a href="http://www.kb.dk/subject2112/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Fotografarkiver</h:a> / <h:a
href="http://www.kb.dk/subject668/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Türck, Sven</h:a><h:a href="http://www.kb.dk/subject668/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Türck, Sven</h:a> / <h:a
href="http://www.kb.dk/subject672/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">topografi, identificeret</h:a><h:a href="http://www.kb.dk/subject672/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">topografi, identificeret</h:a> / <h:a
href="http://www.kb.dk/subject673/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Danmark</h:a><h:a href="http://www.kb.dk/subject673/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">Danmark</h:a></h:div></md:extension><md:extension><!--
node=1331--><h:div xmlns:h="http://www.w3.org/1999/xhtml" xlink:href="#1331"><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Home</h:a><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Hjem</h:a> / <h:a
href="http://www.kb.dk/subject2108/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Billeder</h:a><h:a href="http://www.kb.dk/subject2108/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Billeder</h:a> / <h:a
href="http://www.kb.dk/subject2109/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Samlinger</h:a><h:a href="http://www.kb.dk/subject2109/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Samlinger</h:a> / <h:a
href="http://www.kb.dk/subject2112/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Fotografarkiver</h:a><h:a href="http://www.kb.dk/subject2112/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Fotografarkiver</h:a> / <h:a
href="http://www.kb.dk/subject668/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Türck, Sven</h:a><h:a href="http://www.kb.dk/subject668/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Türck, Sven</h:a> / <h:a
href="http://www.kb.dk/subject1331/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">etaterne</h:a><h:a href="http://www.kb.dk/subject1331/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">etaterne</h:a></h:div></md:extension><md:extension><!--
node=1536--><h:div xmlns:h="http://www.w3.org/1999/xhtml" xlink:href="#1536"><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Home</h:a><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Hjem</h:a> / <h:a
href="http://www.kb.dk/subject2108/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Billeder</h:a><h:a href="http://www.kb.dk/subject2108/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Billeder</h:a> / <h:a
href="http://www.kb.dk/subject2109/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Samlinger</h:a><h:a href="http://www.kb.dk/subject2109/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Samlinger</h:a> / <h:a
href="http://www.kb.dk/subject2112/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Fotografarkiver</h:a><h:a href="http://www.kb.dk/subject2112/da/"

xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Fotografarkiver</h:a> / <h:a
href="http://www.kb.dk/subject668/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Türck, Sven</h:a><h:a href="http://www.kb.dk/subject668/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Türck, Sven</h:a> / <h:a
href="http://www.kb.dk/subject1331/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">etaterne</h:a><h:a href="http://www.kb.dk/subject1331/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">etaterne</h:a> / <h:a
href="http://www.kb.dk/subject1536/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">redningsvæsen</h:a><h:a href="http://www.kb.dk/subject1536/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">redningsvæsen</h:a></h:div></md:extension><md:extension><!--
node=1724--><h:div xmlns:h="http://www.w3.org/1999/xhtml" xlink:href="#1724"><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/en/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="en">Home</h:a><h:a
href="http://www.kb.dk/editions/any/2009/jul/editions/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Hjem</h:a> / <h:a
href="http://www.kb.dk/subject2108/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Billeder</h:a><h:a href="http://www.kb.dk/subject2108/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Billeder</h:a> / <h:a
href="http://www.kb.dk/subject2109/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Samlinger</h:a><h:a href="http://www.kb.dk/subject2109/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Samlinger</h:a> / <h:a
href="http://www.kb.dk/subject2112/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Fotografarkiver</h:a><h:a href="http://www.kb.dk/subject2112/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Fotografarkiver</h:a> / <h:a
href="http://www.kb.dk/subject668/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">Türck, Sven</h:a><h:a href="http://www.kb.dk/subject668/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">Türck, Sven</h:a> / <h:a
href="http://www.kb.dk/subject693/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">transportmidler</h:a><h:a href="http://www.kb.dk/subject693/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">transportmidler</h:a> / <h:a
href="http://www.kb.dk/subject931/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">både</h:a><h:a href="http://www.kb.dk/subject931/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="da">både</h:a> / <h:a
href="http://www.kb.dk/subject1724/en/" xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="en">gummibåde</h:a><h:a href="http://www.kb.dk/subject1724/da/"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang="da">gummibåde</h:a></h:div></md:extension><md:typeOfResource>still
image</md:typeOfResource><md:physicalDescription xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang=""><md:form type="technique">Negativ</md:form></md:physicalDescription><md:subject
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xml:lang=""><md:geographic>Danmark</md:geographic></md:subject><md:accessCondition
xmlns:xml="http://www.w3.org/XML/1998/namespace" xml:lang="">Billedet er muligvis beskyttet af loven om
ophavsret</md:accessCondition><relatedItem xmlns="http://www.loc.gov/mods/v3"/><md:identifier type="uri"
xml:lang="da">http://www.kb.dk//images/billed/2010/okt/billeder/da</md:identifier><md:identifier type="uri"
xml:lang="en">http://www.kb.dk//images/billed/2010/okt/billeder/en</md:identifier>
</md:mods></metadata></record>

Appendix B: Master File, XSLT, and Resultant XML

```
<?xml version="1.0" encoding="UTF-8"?>
<files>
   <file>oai.xml</file>
   <file>oai01.xml</file>
   <file>oai02.xml</file>
   <file>oai03.xml</file>
   <file>oai04.xml</file>
   <file>oai05.xml</file>
   <file>oai06.xml</file>
   <file>oai07.xml</file>
   <file>oai08.xml</file>
   <file>oai09.xml</file>
   <file>oai10.xml</file>
   <file>oai11.xml</file>
   <file>oai12.xml</file>
   <file>oai13.xml</file>
   <file>oai14.xml</file>
   <file>oai15.xml</file>
   <file>oai16.xml</file>
   <file>oai17.xml</file>
   <file>oai18.xml</file>
   <file>oai19.xml</file>
   <file>oai20.xml</file>
   <file>oai21.xml</file>
   <file>oai22.xml</file>
   <file>oai23.xml</file>
   <file>oai24.xml</file>
   <file>oai25.xml</file>
   <file>oai26.xml</file>
   <file>oai27.xml</file>
   <file>oai28.xml</file>
   <file>oai29.xml</file>
   <file>oai30.xml</file>
   <file>oai31.xml</file>
   <file>oai32.xml</file>
   <file>oai33.xml</file>
   <file>oai34.xml</file>
   <file>oai35.xml</file>
   <file>oai36.xml</file>
   <file>oai37.xml</file>
</files>
```

XSLT:

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"  version="1.0"
   xmlns:md="http://www.loc.gov/mods/v3"
   xmlns:h="http://www.w3.org/1999/xhtml"
   xmlns:exsl="http://exslt.org/common"
   extension-element-prefixes="exsl"
   exclude-result-prefixes="md h xsl">

<!-- Corrects spacing in result document -->
   <xsl:strip-space elements="*"/>
   <xsl:output method="xml" indent="yes"/>

 <!-- Allows contents of 'file' element to be recognized as documents. -->
   <xsl:template match="file">
         <xsl:apply-templates select="document(.)"/>
   </xsl:template>

   <xsl:template match="metadata">
      <xsl:apply-templates/>
   </xsl:template>

   <xsl:template match="md:mods">
    <!-- If statement allows this template to apply only if the record creator is 'Türck, Sven (1897-1954) fotograf'
-->
      <xsl:if test="md:name/md:role/md:roleTerm='creator' and md:name/md:namePart='Türck, Sven (1897-
1954) fotograf'">
    <!-- The OAI metadata does not include the URI for the image, so these variables use the record identifier to
build the URL for another resource that does include this information. -->
         <xsl:variable  name="mods_uri"><xsl:value-of select="concat('http://www.kb.dk/cop/syndication',
md:recordInfo/md:recordIdentifier, '?format=mods&amp;lang=en')"/></xsl:variable>
         <xsl:variable name="mods_doc" select="document($mods_uri)"/>
   <!-- Extracts the author, record identifier, title, subjects, and categories from the record. -->
         <record>
            <author>
               <xsl:value-of select="md:name/md:namePart"/>
            </author>
            <recordIdentifier>
               <xsl:value-of select="md:recordInfo/md:recordIdentifier"/>
            </recordIdentifier>
            <title>
               <xsl:value-of select="md:titleInfo/md:title"/>
            </title>
            <xsl:for-each select="md:subject">
               <subject>
                  <xsl:value-of select="descendant::text()"/>
               </subject>
            </xsl:for-each>
            <xsl:for-each select="md:extension/h:div">
               <xsl:if test="not(h:a[last()]='Türck, Sven')">
                  <category>
                     <xsl:value-of select="h:a[last()]"/>
```

```
            </category>
          </xsl:if>
        </xsl:for-each>

 <!-- Uses the variables from above to reference an online XML file, and locate the image URL. -->
        <image>
          <xsl:value-of select="exsl:node-set($mods_doc)//md:identifier[@displayLabel='image']"/>
        </image>
      </record>
    </xsl:if>
  </xsl:template>

<!-- Removes extraneous text -->
  <xsl:template match="text()"/>
</xsl:stylesheet>
```

XML:

```
<?xml version="1.0" encoding="utf-8"?>
<records>
<record>
  <author>Türck, Sven (1897-1954) fotograf</author>
  <recordIdentifier>/images/billed/2010/okt/billeder/object72513</recordIdentifier>
  <title>Kystlinie</title>
  <subject>Danmark</subject>
  <category>Danmark</category>
  <category>topografi, uidentificeret</category>
  <category>strande</category>
  <category>landskabsfotografi</category>
  <image>http://www.kb.dk/imageService/online_master_arkiv_3/non-
archival/samlingsbilleder/turck/turck_02038.jpg</image>
</record>
<record>
  <author>Türck, Sven (1897-1954) fotograf</author>
  <recordIdentifier>/images/billed/2010/okt/billeder/object71437</recordIdentifier>
  <title>Kystparti med høfder</title>
  <subject>Danmark</subject>
  <category>Danmark</category>
  <category>topografi, uidentificeret</category>
  <category>kyster</category>
  <category>strande</category>
  <image>http://www.kb.dk/imageService/online_master_arkiv_3/non-
archival/samlingsbilleder/turck/turck_02139.jpg</image>
</record>
```

Appendix C: XSLT to create HTML table

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">

<!-- Puts Turck metadata into an HTML table. -->
<xsl:output method="html" indent="yes"/>
   <xsl:template match="record">
      <tr>
         <td><xsl:value-of select="image"/></td>
         <td>
          <xsl:for-each select="subject">
             <xsl:value-of select="."/><xsl:text>; </xsl:text>
          </xsl:for-each>
         </td>
          <td>
             <xsl:for-each select="category">
                <xsl:value-of select="."/><xsl:text>; </xsl:text>
             </xsl:for-each>
          </td>
          <td><xsl:value-of select="title"/></td>
          <td><xsl:value-of select="recordIdentifier"/></td>
      </tr>
   </xsl:template>

   <xsl:template match="/">
      <html>
      <table>
         <xsl:apply-templates/>
      </table>
      </html>
   </xsl:template>
</xsl:stylesheet>
```

http://www.kb.dk/imageService/online_master_arkiv_3/non-archival/samlingsbilleder/turck/turck_62984.jpg	Danmark;	børn; Danmark; drenge; beklædning; brændeglas; sweatre;	Dreng med brændeglas	/images/billed/2010/okt/billeder/object63671
http://www.kb.dk/imageService/online_master_arkiv_3/non-archival/samlingsbilleder/turck/turck_06051.jpg		Danmark; vasketøj; tørresnor; baggårde; ;	Baggård med vasketøj	/images/billed/2010/okt/billeder/object61646
http://www.kb.dk/imageService/online_master_arkiv/non-archival/samlingsbilleder/turck/turck_61294.jpg	Danmark;	Danmark; topografi, uidentificeret; byrum; provinsbyer; bindingsværk; byhuse;	Provinsby	/images/billed/2010/okt/billeder/object77301
http://www.kb.dk/imageService/online_master_arkiv_3/non-archival/samlingsbilleder/turck/turck_17898.jpg	Danmark, Knivsbjerg;	Danmark; landskabsfotografi; Sønderjylland; Knivsbjerg; ;	Knivsbjerg-tårnet eller Bismarcktårnet	/images/billed/2010/okt/billeder/object64921

Appendix D: XSLT to remove duplicates

All game tags (in Excel XML)

…

```
<Row>
   <Data
ss:Type="String">http://www.kb.dk/imageService/w330/h330/o
nline_master_arkiv_3/non-
archival/samlingsbilleder/turck/turck_06002.jpg</Data>
   <Data ss:Type="String">plante</Data>
   <Data ss:Type="String">plante</Data>
   <Data ss:Type="String">træ</Data>
   <Data ss:Type="String">blomst</Data>
   <Data ss:Type="String">blade</Data>
   <Data ss:Type="String">klokker</Data>
   <Data ss:Type="String">tobaksplante</Data>
   <Data ss:Type="String">blomster</Data>
   <Data ss:Type="String">blomster</Data>
   <Data ss:Type="String">tromptengle</Data>
   <Data ss:Type="String">blomst</Data>
   <Data ss:Type="String">træ</Data>
   <Data ss:Type="String">natur</Data>
   <Data ss:Type="String">blomster</Data>
  </Row>
  <Row>
   <Data
ss:Type="String">http://www.kb.dk/imageService/w330/h330/o
nline_master_arkiv_3/non-
archival/samlingsbilleder/turck/turck_01201.jpg</Data>
   <Data ss:Type="String">snak</Data>
   <Data ss:Type="String">publikum</Data>
   <Data ss:Type="String">arrangement</Data>
   <Data ss:Type="String">kvinder</Data>
   <Data ss:Type="String">park</Data>
   <Data ss:Type="String">græs</Data>
   <Data ss:Type="String">dame</Data>
   <Data ss:Type="String">mennesker</Data>
   <Data ss:Type="String">mennesker</Data>
   <Data ss:Type="String">folkemængde</Data>
   <Data ss:Type="String">mennesker</Data>
   <Data ss:Type="String">park</Data>
   <Data ss:Type="String">græsplæne</Data>
   <Data ss:Type="String">park</Data>
  </Row>
```

Validated tags (in HTML table)

…

```
<tr>
<td>http://www.kb.dk/imageService/w330/h330
/online_master_arkiv_3/non-
archival/samlingsbilleder/turck/turck_06002.jpg
</td>
<td>plante; blomster; blomst; blade; tr&aelig;;
</td></tr>
  <tr>
<td>http://www.kb.dk/imageService/w330/h330
/online_master_arkiv_3/non-
archival/samlingsbilleder/turck/turck_01201.jpg
</td>
    <td>mennesker; park; </td></tr>

…
```

XSLT

Find Duplicates

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
   <xsl:strip-space elements="*"/>
```

```
  <xsl:output method="xml" indent="yes"/>

<!-- Checks whether a previous data field has the same value as the current one, and, if so, includes it. -->
  <xsl:template match="Worksheet/Table/Row">
    <tr>
      <td><xsl:value-of select="Data[1]"/></td>
      <xsl:for-each select="Data">
      <xsl:if test="preceding-sibling::Data[.=string(current())]">
      <td><xsl:value-of select="."/></td>
      </xsl:if>
      </xsl:for-each>
    </tr>
  </xsl:template>

<xsl:template match="text()"></xsl:template>
  <xsl:template match="/">
    <table>
      <xsl:apply-templates/>
    </table>
  </xsl:template>
</xsl:stylesheet>
```

Eliminate Duplicates (for tags that appear 3+ times)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
  <xsl:strip-space elements="*"/>
  <xsl:output method="html" indent="yes"/>

<!-- Only includes the first occurence of each tag value. -->
  <xsl:template match="tr">
    <xsl:variable name="http" select="td[1]"></xsl:variable>
    <tr>
      <td><xsl:value-of select="td[1]"/></td>
     <td> <xsl:for-each select="td">
        <xsl:if test="not(.=$http)">
       <xsl:if test="not(preceding-sibling::td[.=string(current())])">
          <xsl:value-of select="."/><xsl:text>; </xsl:text>
       </xsl:if></xsl:if>
      </xsl:for-each>  </td>
    </tr>
  </xsl:template>

<xsl:template match="text()"></xsl:template>
  <xsl:template match="/">
    <table>
      <xsl:apply-templates/>
    </table>
  </xsl:template>
</xsl:stylesheet>
```

Appendix E: Sample of collected data in Excel after standardizing

image_id (for validation)	Free tags	Number of Free tags	2VTags	Number of 2Vtags	3Vtags	Number of 3Vtags	Descriptors	Number of Descriptors
http://www.kb.dk/images/billed/2010/okt/billeder/da/object61569/	snak;publikum;arrangement;kvinder;park;græs;dame;mennesker;tilhørere;optræden;mennesker;folkemængde;mennesker;park;græsplæne;park	16	mennesker; park;	2	mennesker;park;	2	balloner; ferie og fritid; folkeliv ; forlystelseslivet i Danmark; klapvogne; plakatsøjler; ;	9
http://www.kb.dk/images/billed/2010/okt/billeder/da/object61571/	fanødragt;fanøkone;dame;kone;kjole;fannik;dame;kop;kjole;kone;amagerkone;gammel;dame;fanø	14	dame; kjole; kone;	3	dame;	1	Danmark; egnsdragter; Fanø; ;	2
http://www.kb.dk/images/billed/2010/okt/billeder/da/object61577/	gammel;hytte;have;hus;træ;hytte;træhus;græstag;græs;træhus;sommerhus;græs;sommerhus;græstag;sommerhus;hegn;fritidshus;træ	18	hytte; træhus; græs; sommerhus; græstag; træ;	6	sommerhus;	1	Danmark; arkitektur; sommerhuse; topografi, uidentificeret; stråtag; ;	3
http://www.kb.dk/images/billed/2010/okt/billeder/da/object61579/	stråtag;strandhus;TRUE;hav;stråtag;vand;strand;øer;hav;strandhus;vindue;udsigt;kyst;stråtækt;strandbo;hus	16	stråtag; hav; strandhus;	3		0	Danmark; arkitektur; sommerhuse; stråtag; ;	3
http://www.kb.dk/images/billed/2010/okt/billeder/da/object61580/	sommerhus;sommerhus;hus;dannebrog;flag;sommerhus;dannebrog,_strand,_sommer;stråtag;strand;kyst;ved_havet;strand;dannebrog;strand;flag;vesterhavet;dannebrog;flag;hus	19	sommerhus; strand; dannebrog; flag; hus;	5	sommerhus;dannebrog;flag;strand;	4	Danmark; arkitektur; sommerhuse; topografi, uidentificeret; stråtag; ;	3