



PÅ

flickr LOVES YOU
TM

- En analyse af tagsene tilføjet Library of Congress' billeder på Flickr

Bachelorprojekt af

Mia Nørhave Bisgaard

Danmarks Biblioteksskole, Aalborg 2008

Vejleder: Brian Kirkegaard

Abstract: I nærværende projekt analyseres 1.277 tags, der er tilføjet 30 af de fotografier, som Library of Congress har lagt ud på billeddelingssiden Flickr. Formålet med analysen er at kunne vurdere om en umiddelbar implementering af tagsene i Library of Congress' billeddatabase vil være fordelagtig. Dette afhænger af hvilke typer tagsene er af og om de vil være brugbare og værdiberigende for databasen, hvorfor dette undersøges i projektet. Resultaterne fra analysen af tagsene viste at 80 % af dem er udtryk for billedernes emner. Størstedelen er tags, der beskriver generelle personer, dyr eller ting. Dernæst beskriver hovedparten af tagsene specifikke steder, billedernes aboutness og specifikke personer, dyr eller ting. Ved at sammenligne resultaterne fra undersøgelsen og den relaterede litteratur vurderes tagsene til overvejende at være både brugbare og værdiberigende for databasen, hvis LOC er villige til at acceptere risikoen for, at få af tagsene kan indeholde fejlagtige informationer.

Indholdsfortegnelse

1. INDLEDNING	3
1.1 Problemstilling	3
1.2 Problemformulering	4
1.3 Læsevejledning	5
2. RELATERET LITTERATUR	5
2.1 Billedindeksring	5
2.2 Billedsøgeadfærd	7
2.3 Tagging	9
2.4 Flickr	11
3. METODE	13
3.1 Dataindsamling	13
3.2 Kategorisering af tags	13
4. ANALYSE OG RESULTATER.....	17
5. DISKUSSION	20
5.1 Diskussion af analysens resultater	20
5.2 Diskussion af metodevalget.....	24
6. KONKLUSION	25
7. PERSPEKTIVERING	26
8. LITTERATURLISTE.....	27
BILAG 1: De udvalgte billeder med titel, tags og URL	
BILAG 2: Cd-rom med projektet, bilag 1 samt kategoriseringen af tagsene	

The Library of Congress på Flickr

- en analyse af tagsene tilføjet Library of Congress' billeder på Flickr

1. Indledning

1.1 Problemstilling

Det såkaldte "web 2.0" vinder større og større udbredelse. Udtrykket dækker over en række forskellige sociale teknologier, der blandt andet har det til fælles, at brugerne selv skaber og personaliserer indholdet. (Farkas, 2007) Eksempler på disse teknologier er blogs, wikis, billeddelingssider som Flickr (<http://www.flickr.com>), sociale netværkssider som MySpace (<http://www.myspace.com>) og Facebook (<http://www.facebook.com>) samt tjenester som Del.icio.us (<http://del.icio.us>), hvor det er muligt at organisere bookmarks. Mange af de sociale teknologier er baseret på filosofien "the wisdom of crowds": flere hoveder tænker bedre end ét. Disse teknologier skaber således grundlag for en enorm vidensdeling, da de giver mulighed for at samarbejde, samtale og opbygge fællesskaber på nettet. Ved at benytte teknologierne kan der derfor ofte trækkes på flere hundrede eller tusinde menneskers viden. En viden det ville være meget ressourcekrævende eller direkte umulig at skaffe på anden måde. (Farkas, 2007)

For BDI-verdenen betyder udbredelsen af "web 2.0" udfordringer i forhold til, hvordan de sociale teknologier med fordel kan benyttes til at forbedre eksempelvis organisering og genfindning.

Library of Congress (LOC) har taget en af disse "web 2.0"-komponenter til sig ved at lægge over 3.500 billeder¹ ud på Flickr og opfordre sidens brugere til at tagge og kommentere dem. Billederne er fra LOC's "Prints & Photographs Online Catalog", der giver adgang til mere end 50 % af afdelingens samling af billeder (Library of Congress, 2007). Mange af billederne mangler centrale informationer om motiverne, såsom hvor billedet er taget og hvem, som er på billedet. Formålet med LOC's Flickr-projekt er at forbedre fotografiernes metadata med henblik på at sikre den bedst mulige information om fotografierne samt en til stadighed bedre adgang til samlingen. (Raymond, 2008a) Flickr-projektet har kørt siden januar 2008 og allerede i marts 2008 var 68 af de bibliografiske poster blevet modificeret af indeksererne

¹ Antallet af billeder 08.05.08 er 3.515

med hjælp fra brugerne på Flickr (Raymond, 2008b)². Indtil videre har ændringerne primært været baseret på de kommentarer, som brugerne har tilføjet de enkelte billeder, mens tagsene endnu ikke direkte er blevet udnyttet.

1.2 Problemformulering

På baggrund af ovenstående findes det interessant at undersøge de tags, der er tilføjet LOC's billeder på Flickr nærmere i dette projekt. Formålet med undersøgelsen er at vurdere, hvorvidt en umiddelbar implementering af tagsene i LOC's billeddatabase kunne skabe en endnu bedre adgang til billederne. Problemformuleringen lyder derfor som nedenstående:

Vil en umiddelbar implementering af tagsene tilføjet Library of Congress' billeder på Flickr i "Prints & Photographs Online Catalog" være fordelagtig?

Dette er interessant at undersøge, da en implementering af allerede eksisterende metadata vil være en meget ressourcebesparende måde at tilføje flere emneindgange på. Det er også interessant, idet tags ikke følger principperne for emneord i traditionel indeksering, hvorfor det kunne være spændende at undersøge, om de kan bruges som sådanne alligevel.

Måden, hvorpå det ønskes at besvare ovenstående problemformulering er ved hjælp af følgende to underspørgsmål:

- 1) **Hvilke typer af tags er tilføjet billederne fra Library of Congress på Flickr?**
- 2) **Er tagsene brugbare og værdiberigende for databasen?**

Underspørgsmål 1 er relevant, idet en eventuel implementering kun vil være fordelagtig, hvis tagsene er af en type, som databasens brugere vil søge på. Ved at undersøge typen af tags og sammenligne disse med resultater fra undersøgelser af brugernes søgeadfærd, er formålet at vurdere, hvorvidt tagsene vil bidrage med relevante søgeindgange i databasen. Type betyder i denne undersøgelse både om tagsene eksempelvis er emneord eller bibliografiske data og hvilken slags emneord eller bibliografiske data, der er tale om.

Det vil også være relevant at undersøge om tagsene er brugbare og værdiberigende for databasen, hvorfor underspørgsmål 2 er medtaget. Ifølge Politikens Nudansk Ordbog betyder *brugbar*: "som kan bruges til et bestemt formål i en bestemt sammenhæng" (Becker-

² Se http://www.flickr.com/photos/library_of_congress/2179058148/ og http://www.flickr.com/photos/library_of_congress/2179042924/ for eksempler.

Christensen, 1999), hvilket også er den betydning ordet er tiltænkt i denne undersøgelse. Her er formålet genfinding og sammenhængen er LOC's billeddatabase. Hvis tagsene ikke er brugbare som søgeindgange til genfinding, vil en implementering ikke være fordelagtig. *Værdiberigende* betyder i denne sammenhæng, om tagsene tilføjer basen relevante, nye søgeindgange. Hvis dette ikke er tilfældet, vil det ikke ressourcemæssigt kunne betale sig at implementere tagsene.

1.3 Læsevejledning

Opgaven indledes med et afsnit om den relaterede litteratur. Denne danner teoretisk basis for udviklingen af metoden i det efterfølgende afsnit. Metodeafsnittet behandler dels dataindsamlingen og dels udviklingen af et kategoriseringsskema til analyse af tagsene. Efter metodeafsnittet følger den egentlige analyse og en præsentation af resultaterne, hvorefter disse diskuteres i henhold til både undersøgelsesspørgsmålene og den relaterede litteratur i det derpå følgende afsnit. I afsnit 6 konkluderes der på undersøgelsen med en besvarelse af problemformuleringen, hvorefter der endeligt følger en perspektivering med forslag til videre undersøgelser.

2. Relateret litteratur

For at kunne udvikle en metode til besvarelse af problemformulering gennemgås den relaterede litteratur i dette afsnit. Først belyses det i afsnit 2.1, hvilke udfordringer billedindeksering giver i forhold til indekseringen af andre typer af dokumenter samt hvilke metoder og teorier, der kan benyttes til emneindeksering i billedkontekster. Derefter ses der på brugeres søgeadfærd i billeddatabaser, der er sammenlignelige med LOC's. I afsnit 2.3 fokuseres der på fænomenet tagging samt de fordele og ulemper denne form for tilføjelse af metadata giver. Endeligt ses der nærmere på siden Flickr som taggingsystem.

2.1 Billedindeksering

Billeder har bredere og mindre veldefinerede emneindgange end tekstdokumenter, idet de kan beskrives ud fra en lang række faktorer som eksempelvis objekter i motivet, baggrund, metode, indhold, temaer og endda følelser (Choi & Rasmussen, 2003). Hvor der i emneindekseringen af tekstdokumenter såsom monografier og tidsskriftartikler som regel er hjælp at hente i eksempelvis titel, indholdsfortegnelse og abstract (Library og Congress,

1997), er det anderledes med visuelle materialer. Visuel information er kompleks og lader sig ikke direkte oversætte til sprog (Svenonius, 1994). Her kan det samme billede betyde forskellige ting for forskellige mennesker og endda noget forskelligt for det samme menneske på forskellig tid (Shatford, 1986).

I forhold til billedgenfindning er indeksering meget vigtig (Rasmussen, 1997), hvorfor der er udviklet teorier til hjælp med analysen af billeders motiver. Meget billedindekseringsteori er baseret på Panofskys (1972) billedanalyse, der består af tre niveauer af betydning: *pre-iconography* (beskrivelse), *iconography* (analyse) og *iconology* (fortolkning). Dette er på trods af, at teorien oprindeligt blev udviklet med hensigt på analyse af renæssancekunst. Shatford videreudviklede Panofskys teori i 1986 til en facetteret klassifikation af billeders motiver med kategorierne *Generic Of*, *Specific Of* og *About*. Hvad et billede er *af* (*Of*) er konkret og objektivt, mens hvad et billede er *om* (*About*) er mere abstrakt og subjektivt. *Of* inddeler Shatford i *Generic* og *Specific*, hvor *Generic* er generelt/bredt (fx dame) og *Specific* er konkret/specifikt (fx Dronning Margrethe). De tre kategorier kombineres med facetterne: who (hvem), what (hvad), where (hvor) og when (hvornår), som er baseret på Ranganathans (1962) Personality, Matter, Energy, Space og Time. (Shatford, 1986; Shatford Layne 1994)

Facets	Specific of	Generic of	About
1. Who animate and inanimate; concrete objects and beings	Individually named persons, animals, things... (S1)	Kinds of persons, animals, things (G1)	Mythical beings (generic/specific), abstractions manifested or symbolized by objects or beings (A1)
2. What what are the objects and beings doing? (actions events, emotions)	Individually named events (S2)	Actions, conditions (G2)	Emotions, abstractions manifested by actions, events (A2)
3. Where locale, site place; geographic, cosmographic, architectural	Individually named geographic location (S3)	Kind of place, geographic or architectural (G3)	Places symbolized (generic/specific), abstractions manifested by locale (A3)
4. When time; linear or cyclical	Linear time; dates or periods (S4)	Cyclical time; seasons, time of day (G4)	Emotions or abstractions symbolized by or manifested by time (A4)

Tabel 1: Den primære del af Shatfords (1986, s. 46) facetterede klassifikation af billeders emner.

Hvis billederne analyseres manuelt og tildeles emneord fra for eksempel en kontrolleret liste eller ved hjælp af naturligt sprog, benyttes der "concept-based" indeksering (Rasmussen, 1997). Indekseringen af fotografierne i LOC's database er af denne type, idet deres indekserer analyserer billederne og tildeler dem emneord efter deres egen *Theaurus for Graphic*

Materials, *Library of Congress Subject Headings* og *Library of Congress Name Authority File* ud fra et sæt veldefinerede retningslinier (Library of Congress, 1995). Tildelingen af emneord på denne måde vil altid være subjektiv, hvorfor det kan være svært at opnå konsistens (Rasmussen, 1997). Derudover er det en dyr og tidskrævende metode.

Den helt traditionelle indeksering kritiseres også for at hvile på en opfattelse af, at dokumenternes *aboutness* ikke ændrer sig og at den trænede ekspert (indeksøren) er i stand til at afkode dokumentet, hvorpå denne så vil blive universelt accepteret som udgørende betydningen af dokumentet. (Rafferty & Hilderly, 2007)

Modsætningen til ”concept-based” indeksering er ”content-based”, hvor indeksering foretages af en computer. Denne metode har dog også sine begrænsninger. (Enser & Sandom 2006; Rasmussen, 1997) Der er derfor ingen af metoderne, som er optimale.

I det følgende delafsnit ses der nærmere på, hvilken type af emneord brugere af billeddatabaser typisk benytter i deres søgninger.

2.2 Billedsøgeadfærd

Det har ikke været muligt at finde specifikke undersøgelser af søgeadfærden i LOC’s billeddatabase. Der eksisterer dog en del undersøgelser af søgeadfærden i lignende billeddatabaser, som denne litteraturgennemgang vil tage udgangspunkt i. Helt konkret er der udvalgt undersøgelser af Armitage og Enser (1997), Choi og Rasmussen (2003), Collins (1998), Enser (1993) samt van Hooland (2006). Disse er udvalgt, da de alle tager udgangspunkt i tilsvarende databaser, hvorfor de undersøgte brugere må formodes at have en søgeadfærd, der kan sammenlignes med brugerne af LOC’s billeddatabase. Derudover benytter Armitage og Enser (1997), Choi og Rasmussen (2003), van Hooland og til dels også Collins (1998) Shatfords (1986) analyseskema til analysen af søgeformuleringerne, hvilket gør sammenligningen med typen af tags senere i opgaven lettere og mere objektiv.

Armitage og Enser (1997) undersøger søgeadfærden i syv forskellige bibliotekers databaser. I denne opgave er der valgt kun at medtage de to, som vurderes at minde mest om LOC i forhold til samlingens indhold og de formodede brugere.

Enser (1993) benytter sin egen teori, hvor unique er specifik og non-unique er generel. De to kategorier kan specificeres med ”refinements” som relaterer til tid, sted, handling/bevægelse, begivenhed eller tekniske specifikationer. Enser (1993) er medtaget, da resultaterne fra

undersøgelsen stadig er sammenlignelig med de førnævnte og da Enser's undersøgelser generelt er anerkendte og citeret indenfor dette område.

De forskellige undersøgelser og deres resultater er sammenfattet i tabellen nedenfor.

Forfattere	Samling	Analyse baseret på		Resultater
		Population	Teori	
Enser (1993)	The Hulton Deutsch Collection Limited	2722 forespørgsler fra brugerne	Egen	Non-unique: ca. 6% Non-unique, refined: ca. 25% Unique: ca. 42% Unique, refined: ca. 26%
Armitage & Enser (1997)	Glasgow Mitchell Library	170 forespørgsler fra brugerne	Shatford (1986)	S1: 33,9% ³ S2: 0,6% S3: 50,6% S4: 13,1% G1: 46,4% G2: 4,8% G3: 15,5% G4: 0% (A1, A2, A3 og A4 ikke repræsenteret)
	Birmingham Central Library	294 forespørgsler fra brugerne		S1: 39,9% S2: 2,5% S3: 54% S4: 24,5% G1: 14,7% G2: 7,4% G3: 1,2% G4: 0% (A1, A2, A3 og A4 ikke repræsenteret)
Collins (1998)	North Carolina Collection (University of North Carolina) og North Carolina State Archives (Raleigh)	187 forespørgsler fra brugerne	Egen, baseret på blandt andet Panofsky (1972) og Shatford (1986)	Generic: persons: 19%, objects/things: 30%, activities: 16% Specific: personal name: 17%, organization name: 18% Time: decade: 35% Place: city: 19%
Choi & Rasmussen (2003)	Library of Congress' "American Memory" foto arkiv	Forespørgsler fra 38 undervisere og studerende i amerikansk historie	Shatford (1986)	S1: 10,3%, S2: 3,2%, S3: 7,6%, S4: 5,4% G1: 23,8%, G2: 25,4%, G3: 15,7% G4: 0% A1: 2,7%, A2: 4,9%, A3: 1,1%, A4: 0%
Van Hooland (2006)	National Archives of the Netherlands	384 forespørgsler fra brugerne	Shatford (1986)	S1: 17,5%, S2: 5,5%, S3: 57%, S4: 2,5% G1 9%, G2: 8,5% (G3, G4, A1, A2, A3 og A4 ikke repræsenteret)

Tabel 2: Undersøgelser af brugeres søgeadfærd i billeddatabaser

Resultaterne fra undersøgelserne er meget ens, hvilket underbygger antagelsen om, at brugerne af LOC's billeddatabase kan have en tilsvarende søgeadfærd.

Som det kan aflæses af tabellen, har brugerne i undersøgelserne hovedsagligt søgt på specifikke geografiske steder (S3) undtagen i undersøgelsen af Choi og Rasmussen (2003), hvor størstedelen af søgningerne var på generelle handlinger (G2). Dette er ellers en type, der ikke er benyttet så meget i de andre undersøgelser. G1 og S1, der er henholdsvis generelle og

³ I undersøgelsen blev nogle af forespørgslerne vurderet til at indeholde flere aspekter, hvorfor det samlede procenttal er over 100 % (Armitage & Enser, 1997).

specifikke ting, dyr og personer, er den type af søgetermer, som benyttes næst mest i databaserne. Brugere søger også til vis grad på specifikke datoer og perioder (S4) samt generelle steder (G3). Der er meget få af brugerne som benytter abstrakte termer i deres søgeformuleringer og i nogle af undersøgelserne er det slet ingen. Derudover er G4 (generel tid) slet ikke repræsenteret.

Basen "American Memory", der blev benyttet i Choi og Rasmussens (2003) undersøgelse, indeholder mange af de samme billeder som "Prints & Photographs Online Catalog" gør. Grundet dette må denne base være den mest sammenlignelige. Det er derfor væsentligt at lægge vægt på resultaterne fra denne undersøgelse på trods af, at disse adskiller sig fra resultaterne fra de andre undersøgelser med hensyn til den type af søgetermer, der er mest benyttet. Det er dog stadig vigtigt at inddrage de resterende undersøgelser, idet der i undersøgelsen af Choi og Rasmussen (2003) kun blev analyseret forespørgsler fra historielærere og –studerende. Man må formode at LOC's billeddatabase har en bredere brugergruppe.

I næste delafsnit ses der nærmere på fænomenet tagging samt hvilke fordele og ulemper, der er ved denne form for metadata.

2.3 Tagging

Tagging er når man lader alle – specielt brugerne – tilføje termer (tags) til dokumenter. Det er mest brugbart, når der ingen er i bibliotekarrollen eller der er for meget indhold til, at en enkelt autoritet kan indekserer det hele. (Golder & Huberman, 2005)

Idet det typisk er brugerne, som tilføjer termer til ressourcerne, adskiller tagging sig fra indeksering på en række områder. Selve typen af dokumenter kan være alt fra websider til ideer og ikke kun værker, som det typisk er i en BDI-kontekst. Brugere af de tilføjede termer er taggerne selv og næsten alt tagging har personlig organisering af dokumenter til formål. Tagsene kan dog oftest med fordel bruges af andre. (Golder & Huberman, 2005; Tennis, 2006)

Ifølge Hammond og kollegaer (2005) er tagsene på Flickr udpræget 'egoistiske', idet brugerne i overvejende grad kun tagger deres egne billeder og kun for at lette genfindingen af billederne for dem selv. I en undersøgelse af tags tilføjet billeder på Flickr i en universitetskontekst vurderes over halvdelen af de analyserede tags dog som værende

brugbare for alle brugere (Angus, Thelwall & Stuart 2008). Marlow (2006) argumenterer da også for at motivationen for tagging ligeledes kan være social, idet brugerne ønsker at dele deres ressourcer, tiltrække opmærksomhed, konkurrere eller udtrykke deres mening. Det er i så fald vigtigt at tagsene er brugbare for andre end dem selv.

I modsætning til traditionel indeksering er tagging hverken eksklusiv eller hierarkisk. Det samme dokument kan derfor være tilføjet en bred variation af termer på samme tid og på forskellige specificitetsniveauer. Det betyder, at et dokument eksempelvis kan være identificeret til at handle om Afrika, katte, dyr, geparder osv. på samme tid, hvilket både kan være en fordel og en ulempe i forhold til offentlig genfindning. (Golder & Huberman, 2005; Rafferty & Hilderly, 2007)

I langt størstedelen af taggingsystemerne er der ingen kontrol med homonymer, synonymer, bøjningsformer og stavning, hvilket giver stor risiko for støj i søgninger (Golder & Huberman, 2005; MacGregor & McCulloch, 2006). Derudover benyttes der ofte koder, som kun giver mening for den enkelte tagger og sammensatte ord, der ikke adskilles af mellemrum. Endeligt kan der forekomme ”falsk” brug af tags, som gør genfindingen mindre pålidelig. Det kan eksempelvis være ordet ”bryllup”, som er tilføjet et billede af et eger, fordi billedet er taget ved et bryllup. (Rafferty & Hilderly, 2007) Det er dog muligt at forbedre kvaliteten af tags, hvis der eksempelvis indføres nogle retningslinier for, hvordan formen og typen bør være i et system. (Guy & Tonkin, 2006)

Fordelene ved tagging er blandt andet en reducere i omkostningerne i forhold til traditionel indeksering. Derudover kan tagsene give flere emneindgange, der samtidigt i højere grad reflekterer brugernes ordforråd. Det giver dermed bedre muligheder for serendipity og browsing, som ofte er brugernes søgeteknikker i billeddatabaser. (Frost et. al., 2000; MacGregor & McCulloch, 2006; Mathes, 2004; Rafferty & Hilderly, 2007)

I Golder og Huberman’s (2005) undersøgelse af tagsene tilføjet på Del.icio.us, har de identificeret syv typer af tags:

1. Hvad (eller hvem) dokumentet er om
2. Hvad dokumentet er (artikel, blog, bog)
3. Hvem der ejer dokumentet

4. Specificerende kategorier (underkategorier)
5. Kvaliteter eller karakteristika (sjov, skræmmende, inspirerende)
6. Selvreferende (tags, der begynder med "my" – ex. Mystuff)
7. Organisering af opgave (toread, jobsearch)

Det kan diskuteres, hvorvidt tagging er en form for indeksering eller ej. Ifølge informationsordbogen er indeksering: "Den proces at analysere dokumenter med henblik på udarbejdelse af beskrivende og emnemæssige indførsler til registre" (Andersen, 2006). I tagging analyseres dokumenterne med det formål at tilføje dem emneord eller i hvert fald termer, der kan hjælpe til en senere genfindning. Der ligger dog en stor forskel i, om man tilføjer termer for at lette genfindingen for sig selv eller for alle andre. I tagging er det, som beskrevet ovenfor typisk det første, der er tilfældet.

Shirky (2005) argumenterer for at tagging helt klart er en indekseringsform og ligefrem den eneste, der vil eksistere i fremtiden. Der er dog i den øvrige litteratur bred enighed om, at tagging ikke kan afløse traditionel indeksering, men muligvis kan styrke og forbedre den, hvis den bruges som supplement. (Kipp & Campbell, 2006; Matusiak, 2006; van Hooland, 2006) Hvis tagging benyttes som supplement vil det dog ikke gøre indekseringen mindre subjektiv eller mere konsistent, hvilket var nogle af de problemer, der var i traditionel indeksering – snarere tværtimod. Det ville dog bidrage med de fordele, der blev nævnt tidligere i afsnittet.

I nedenstående delafsnit ses der nærmere på, hvordan tagging benyttes på siden Flickr og hvordan det som taggingsystem adskiller sig fra andre lignende systemer.

2.4 Flickr

Flickr blev introduceret af den canadiske virksomhed Ludicorp i 2004. Det er en side, hvor det er muligt at uploade billeder og film med det formål at opbevare, dele og/eller organisere dem. Derudover er det også muligt at oprette netværk på siden ved at samle sine venner, blive medlem af forskellige grupper, sende beskeder til andre brugere, udvælge favoritbilleder, tilføje tags og kommentarer og meget mere. (Wikipedia bidragere, 2008a)

I litteraturen bliver Flickr ofte skåret over samme kam som sider såsom Del.icio.us, men det er ikke helt retfærdigt. Selvom brugen af tags på mange måder er ens, er der markante forskelle. Marlow med flere (2006) har udviklet en taksonomi over de forskellige

taggingssystemers design, idet de argumenterer for, at forskelle i systemerne har indflydelse på brugen og typen af tags. Dimensioner i taksonomien er blandt andet, hvem der har rettigheder til at tage dokumenterne, hvilken type af ressource, der tagges og om det er muligt at tilføje det samme tag til en ressource flere gange.

I Flickr uploader den enkelte bruger sine egne billeder og har dermed rettigheder over disse. Det er så op til brugeren at bestemme, om andre må tilføje tags og kommentarer til billederne. Marlow og hans kollegaer (2006) analyserede 58 millioner af tagsene på Flickr i deres undersøgelse og fandt derved frem til, at kun en lille delmængde af disse var tilføjet af andre end ejeren. Dette har betydning for typen af tags, idet der kan være forskel på, om et billede tagges af fotografen, deres venner eller komplet fremmede (Marlow et. al., 2006). Ifølge Cox, Clough og Marlow (2008) tilføjer brugerne ikke tags til andres billeder, da det anses som værende uhøfligt og en invasion af privatsfæren. Dette skyldes også, at det ikke er muligt at se, hvem der har tilføjet tagsene.

De tags, der er tilføjet LOC's billeder kan derfor tænkes at adskille sig fra de typiske tags anvendt på Flickr, idet det netop her er alle andre brugere end ejeren selv, som tilføjer tags.

Hvert tag kan kun tilføjes det enkelte billede én gang på Flickr, hvilket står i kontrast til eksempelvis Del.icio.us' taggingssystem, hvor alle har adgang til at tage ressourcerne og hvor der ingen begrænsninger er for brugen af tags. Det har den ulempe, at det ikke er muligt at se, hvad flest mener beskriver motivet bedst. Dermed opnås "the wisdom of crowds" ikke, hvilket ellers er en af fordelene ved tagging. I taggingssystemer som Del.icio.us dannes der ofte konsensus omkring et vis antal tags, der vurderes som dækkende for ressourcen (Kipp & Campbell, 2006).

"Although Flickr is often discussed as part of the tagging phenomenon, the discrete nature of uploaded objects prohibits such a "close knit society" and thus "collaborative tagging" – as distinct from "tagging" – is not made possible" (MacGregor & McCulloch, 2006, s. 295).

Resultaterne fra undersøgelserne omkring både billedindeksering, søgeadfærd i billeddatabaser, taggingfænomenet og siden Flickr vil i det følgende afsnit danne basis for udviklingen af en metode til besvarelse af problemformuleringen.

3. Metode

For at kunne besvare problemformuleringen bliver undersøgelsen udført i to hoveddele: Dataindsamling og kategorisering af tags.

3.1 Dataindsamling

LOC har arrangeret deres billeder på Flickr i to sæt, hvor der er henholdsvis nyhedsbilleder fra 1910'erne og farvebilleder fra 1939-1944. Der bliver til stadighed lagt nye billeder ud, hvorfor der kun er udvalgt billeder fra januar, hvor de første billeder blev lagt ud. Denne udvælgelse er for at sikre, at der er tilføjet flest mulige tags til billederne. Da der er stor forskel på billederne i de to sæt, er der udvalgt 15 billeder fra hver. Antallet vurderes passende som analysegrundlag, idet det giver tilpas mange tags, til at kunne sige noget om typen. Der er udvalgt billeder med en bred variation af motiver såsom landskaber og bybilleder, med og uden mennesker, begivenheder og portrætter osv. for at få et varieret udsnit. Ingen af de valgte billeder har under 20 tags, hvilket igen sker for at sikre et vist analysegrundlag. De udvalgte billeder med titel, tags og URL kan ses i Bilag 1.

3.2 Kategorisering af tags

På baggrund af den relaterede litteratur udarbejdes et skema til kategoriseringen af tagsene. Selve skemaet er hovedsagligt baseret på Shatford (1986), Golder og Huberman (2006), Rafferty og Hidderly (2007), Angus, Thelwall og Stuart (2008) samt min egen erfaring, der er opnået ved at se LOC's billeder på Flickr igennem. Derudover bærer alle overvejelser præg af den relaterede litteratur som hele.

"Kategorisering" er valgt frem for "klassifikation". Dette skyldes, at kategorisering er mindre streng og med mindre klare grænser. Kategorisering er i højere grad baseret på en sammenfatning af lighed end et systematisk, hierarkisk arrangement af materialer. (Mathes, 2004) Dette harmonerer med, at det er tags, der analyseres, og ikke eksempelvis emneord fra en thesaurus.

Skemaet, som kan ses nedenfor, er opdelt i fire hovedgrupper, der relaterer til emneindeksering (A), bibliografiske data (B), sociale tags (C) og fejlkilder (D).

Kategorier		Uddybning	Kilde
A1	Generel person, dyr eller ting	Tag, der identificerer hvad billedet er af på det mest objektive og generelle niveau. Specifik viden om billedet er ikke nødvendig. (fx mand, kat, stol)	Shatford (1986) (G1)
A2	Generelt sted	Tag, der identificerer typen af sted. Specifik viden om billedet er ikke nødvendig (fx landskab, by, jungle, marked)	Shatford (1986) (G2)
A3	Generel begivenhed eller handling.	Tag, der identificerer, hvad objekterne på billedet laver. Specifik viden om billedet er ikke nødvendig (fx fodboldkamp)	Shatford (1986) (G3)
A4	Generel tid	Tag, der identificerer generel tid på dagen eller året (fx vinter, sommer, morgen)	Shatford (1986) (G4)
A5	Specifik person, dyr eller ting	Tag, der kræver baggrundsviden eller specifik viden om motivet på billedet. (fx Theodore Roosevelt)	Shatford (1986) (S1)
A6	Specifikt sted	Tag, der kræver baggrundsviden eller specifik viden om stedet på billedet (fx Eiffeltårnet, Douglas Shoe Factory)	Shatford (1986) (S2)
A7	Specifik begivenhed eller handling	Tag, der kræver baggrundsviden eller specifik viden om begivenheden på billedet (fx 1980 Rosebowl Game)	Shatford (1986) (S3)
A8	Specifik tid	Tag, der referer til årstal, årtier eller lignende tidsangivelse (fx 1942, 1940'erne, fyrerne, marts)	Shatford (1986) (S4)
A9	About	Abstrakte navneord eller tillægsord, der repræsenterer en subjektiv vurdering af hvad billedet forestiller (fx et billede af en gruppe smilende mennesker tagget med "glæde")	Shatford (1986) (A1, A2, A3, A4)
A10	Farver	Subjektive tags, som beskriver fotografiets farver (fx gul, blå, grøn)	LOC's samling på Flickr
B1	Fotograf	Tags, der refererer til den person, som har taget billedet (fx Jack Delano)	LOC's samling på Flickr
B2	Tidligere samling	Tags, der refererer til den samling, billedet oprindeligt var en del af (fx Office of War Information Collection)	LOC's samling på Flickr
B3	Tekniske specifikationer	Tags, der referer til billedets format og andre tekniske specifikationer (ex large format, color, transparencies)	LOC's samling på Flickr
C1	Selvreference	Tags, der referer til taggeren selv (fx mystuff, myphoto)	Golder & Huberman (2005)
C2	Organisation	Tags, der refererer til organisering af opgave – "task-organisation" (fx jobsearch eller toread)	Golder & Huberman (2005)
C3	Sammensatte tags	Tags, der er sammensat uden mellemrum og derfor ikke kan benyttes som søgeterm (fx rosietheriveter)	Angus, Thelwall & Stuart (2008), Rafferty & Hilderly (2007)
C4	Umuligt at bestemme betydning	Tags, der ikke umiddelbart giver nogen mening for andre end taggeren selv (fx koder og slang)	Angus, Thelwall & Stuart (2008), Rafferty & Hilderly (2007)
D1	Stavefejl	Tags, der er stavet forkert og derfor ikke er optimale som søgetermer	Angus, Thelwall & Stuart (2008)
D2	Udenlandske ord	Tags, der er på et andet sprog end engelsk og derfor ikke kan benyttes i LOC's database	Angus, Thelwall & Stuart (2008)
D3	"Falske" ord	Tags, der ikke beskriver motivet på billedet (fx et billede af et eger tagget med ordet "bryllup")	Rafferty & Hilderly (2007)
X	Allerede eksisterende i posten	Tags, der ikke bidrager med ny information, idet termen allerede er til stede i posten	

Tabel 3: Kategoriseringskema benyttet til analyse af tags

A-kategorien er relevant at medtage i analysen af flere årsager. Dels er den væsentlig i forhold til billedindeksering, idet den indeholder de tags, der beskriver billedmotiverne. Derudover er den væsentlig i forhold til at kunne vurdere, om tagsene er værdiberigende for databasen ved at sammenligne resultaterne med undersøgelserne af tilsvarende brugeres søgeadfærd.

Kategorierne er hovedsagligt baseret på Shatford (1986), hvor A1, A2, A3 og A4 referer til *Generic Of*, mens A5, A6, A7 og A8 referer til *Specific Of*. Da undersøgelserne af søgeadfærden (jf. afsnit 2.2) viste, at brugerne sjældent søger på abstrakte begreber, er der valgt kun at medtage delkategorien A9 som dækkende for *About*. Denne kategori indeholder således alle fire facetter. Derudover er kategorien også dækkende for de sociale tags, der beskriver kvaliteter eller karakteristika ved motivet, idet de tags som oftest også vil have form af tillægsord eller abstrakte navneord (Golder & Huberman, 2005). Endeligt er der A10, der indeholder tags, der beskriver motivets farver. Denne er ikke en del af Shatfords (1986) oprindelige klassifikation, men er medtaget i A-kategorien, da den siger noget om billedmotivet.

Kategori B indeholder de tags, der relaterer sig til fotografiernes bibliografiske data. Delkategorierne er lavet løbende, som de udvalgte billeder er blevet analyseret. Tagsene i denne kategori vil typisk ikke være værdiberigende for databasen, idet de oftest allerede er til stede i de enkelte poster.

Kategori C er de sociale tags. Da ”selvrefererende” tags er meget benyttet i taggingsystemer, medtages den som delkategori (C1). Sandsynligheden for, at der findes tags af denne type tilføjet LOC’s billeder, er muligvis lav, idet undersøgelser viste, at Flickr-brugerne ikke benytter sådanne tags til billeder, de ikke selv ejer (jf. afsnit 2.4). De samme argumenter kan benyttes i forhold til C2, der relaterer til organisering af opgaver. Dette er også et tag, der ofte bruges i tagging, men som muligvis ikke er væsentligt benyttet i denne sammenhæng. Dette skyldes både, at brugerne igen ikke er ejerne, men også at der tagges billeder frem for tekst, hvor der ikke er den samme brug for ”task-organisation” (Angus, Thelwall & Stuart, 2008). Begge kategorier medtages dog til enten be- eller afkræftelse af disse antagelser og til vurdering af forskelle eller ligheder mellem tagsene i denne undersøgelse og resultaterne fra andre. Sammensatte tags (C3) og koder (C4) er også typiske i tagging, hvorfor de medtages her. I undersøgelsen af Angus, Thelwall & Stuart (2008) var det eksempelvis 12 % af tagsene, der var sammensatte ord.

Alle tagsene i kategori C er potentielle fejlkilder, idet de med højst sandsynligt ikke er brugbare i LOC's database. De er dog adskilt fra kategori D, der indeholder mere deciderede fejlkilder. Hvis tagsene indeholder stavefejl (D1) eller er på et andet sprog end engelsk (D2), kan de ikke umiddelbart implementeres i databasen og bruges til genfindning. I undersøgelsen af Guy og Tonkin (2006) var 40 % af tagsene fra Flickr enten stavet forkert, på et andet sprog end engelsk, sammensatte ord eller en blanding af sprog. Der er derfor en risiko for, at denne kategori vil være godt repræsenteret i denne undersøgelse. "Falsk" brug af tags (D3) er egentlig relateret til organisering og koder, idet det typisk vil være tags der tilføjes billeder, det ønskes at sammenholde (fx alle de billeder, der er taget til en specifik begivenhed). De bliver dog falske eller forkerte i det øjeblik, de skal være brugbare for andre og ikke er et udtryk for motivet på billedet.

Kategorierne er hovedsagligt eksklusive. Det gælder dog ikke X, der er kategorien til tags, der ikke bidrager med ny information. Hvis alle tagsene tilføjet billederne blot er en gentagelse af de informationer, der allerede er søgbare i posten, vil det være spild af ressourcer at overføre tagsene. Denne kategori er derfor væsentlig at medtage. Tagsene figurerer dog både i X-kategorien og i en af de andre. Derudover kan tags, der tilhører den sociale kategori C3 og fejlkategori D1, samtidigt tilhøre en af de andre kategorier. Dette er valgt på trods af, at formålet med undersøgelsen er at bestemme hvorvidt tagsene umiddelbart vil kunne implementeres i databasen og være brugbare og værdiberigende, idet begge kategorier indeholder tags, der ved få rettelser med stor sandsynlighed vil kunne benyttes. Alle tags tæller dog kun én gang i det samlede antal. Tagget "The Library of Congress" er tilføjet samtlige billeder, hvorfor dette tag ikke tæller med i hverken kategoriseringen eller det samlede antal.

Analysen vil være tilstræbt objektiv, men bærer uundgåeligt præg af mine subjektive vurderinger. Idet kategorierne er eksklusive, er en subjektiv vurdering af om et tag såsom "circus" eksempelvis er en betegnelse for en ting, et sted eller en begivenhed nødvendig. Ligeledes er tagsene af varierende specificitet. Eksempelvis er tagsene "woman", "worker", "riveter" og "rosie the riveter" tags på billede 2, der betegner den samme person, men med fire forskellige grader af specificitet. Sådanne tags er dog alle vurderet til at tilhøre den generelle A1-kategori, idet de ikke betegner en specifik person.

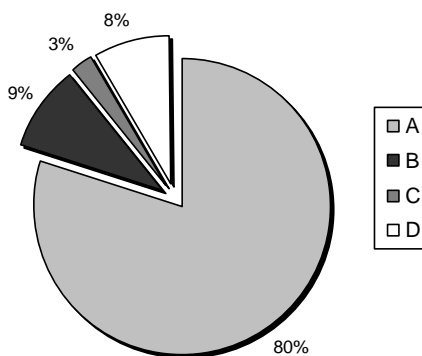
Der skelnes ikke mellem ental og flertal samt enkelt ord og fraser. Dette er ud fra den vurdering af, at alle typer vil være brugbare i basen.

Alle tags analyseres i kontekst med det tilhørende fotografi for bedre at kunne bestemme deres forhold til billedet. Ifølge den komplekse natur af billedfortolkning vil det ikke give mening at se på tagsene i isolation. Det vil eksempelvis også gøre det lettere at bestemme den rette betydning af homonymer.

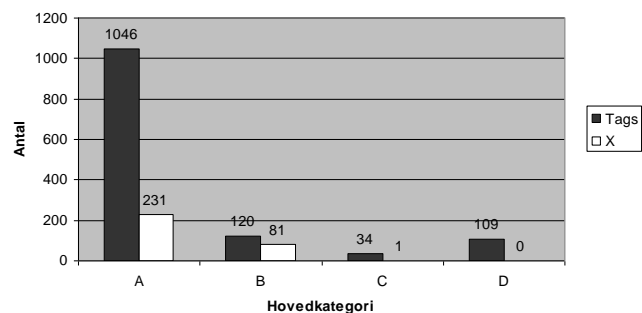
I det følgende afsnit præsenteres selve analysen og dennes resultater.

4. Analyse og resultater

På de 30 udvalgte billeder var der i alt tilføjet 1.277 tags, hvilket i gennemsnit svarer til ca. 43 tags pr. billede. Antallet af tags spænder fra 22 til 72 pr. billede. Der er generelt tilføjet flere tags til farvebillederne end nyhedsbillederne, hvor gennemsnittet er henholdsvis 46 og 39 pr. billede. 75 % af tagsene er nye tags, der ikke allerede er tilgængelige i basen. Hvordan de forskellige tags helt præcist fordeler sig kan ses på den vedlagte cd-rom (bilag 2).

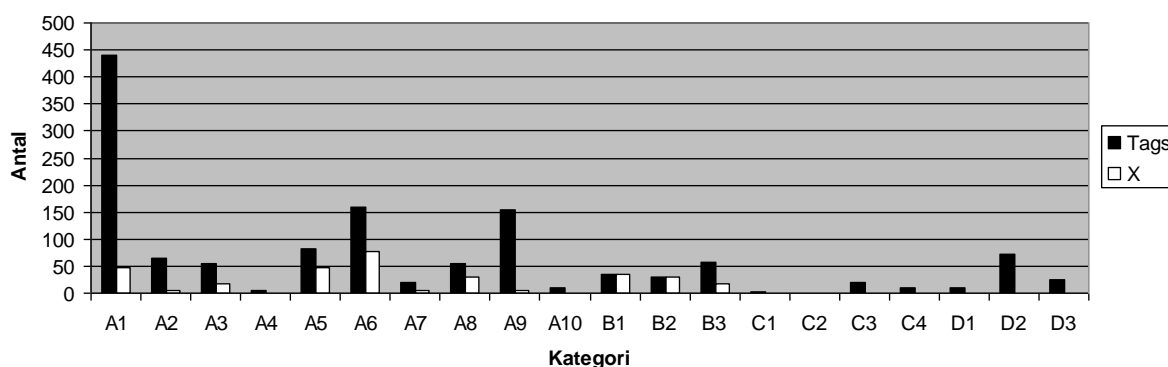


Figur 1: Den procentvise fordeling af tagsene i hovedkategorierne



Figur 2: Antallet af tags i de fire hovedkategorier

Som det kan aflæses af figur 1 og 2 tilhører hovedparten af tagsene kategori A, der er termer, der beskriver motivet på billedet. Af dem var ca. 22 % allerede tilstede i posten. Kun få af de tilføjede tags er enten et udtryk for bibliografiske data (kategori B) eller er deciderede fejlkilder (kategori D). Næsten 70 % af termerne i kategori B var allerede registreret i posten. En meget lille del (ca. 3 %) tilhører kategori C, der er de sociale tags.



Figur 3: Fordelingen af tags i delkategorierne

Når kategorierne bliver udspecificeret som i figur 3, er det tydeligere at se tendenserne i fordelingen. Af de 1.046 tags, der tilhører kategori A, er omkring 42 % af dem et udtryk for en generel person, dyr eller ting (A1). Dette skyldes hovedsagligt at mange af billederne er tagget så ekshaustiv, at nærmest ethvert objekt på motivet er tildelt et ord. Eksempelvis forestiller billede 2 en kvinde, som borer. Udover "woman" og "drill" er billedet dog også tagget med "fingers", "ring" osv. Billederne er generelt tagget meget mere udtømmende, end man ville gøre i traditionel indeksering. Derudover er der benyttet mange synonymer til at beskrive fotografierne. På billede nr. 21 er der eksempelvis tilføjet tagsene "bobbies", "police", "rozzers" og "peelers", der alle er et udtryk for politi eller politimænd. Det er da også kun omkring 10 % af ordene, der allerede er registreret i de nuværende poster. Idet der i kategoriseringen ikke skelnes mellem ental og flertal, er der også flere af termerne som på den måde tæller dobbelt.

Efter A1 er der flest tags, som tilhører A6 (specifikt sted). Under halvdelen af tagsene i denne kategori er nye – dvs. at størstedelen af tagsene allerede er registreret i de pågældende poster. Kategorien indeholder typisk tags, der beskriver geografiske placeringer såsom "Kentucky", "New York", "Brockton" osv. Det er tydeligt, at tagsene ofte er baseret på de eksisterende informationer i posten. Eksempelvis indgår "Coney Island" som en del af titlen på billede nr. 16, der så er tagget med "Coney Island", "New York", "brooklyn", "United States of America" og "canarsie". Det ville have været svært at genkende et sted med så stor præcision, hvis ikke der havde været et fingerpeg tilstede i posten.

Ca. 12 % af det samlede antal tags tilhører kategorien A9 (about). Tagsene er hovedsagligt tillægsord såsom "beautiful", "handsome", "american" osv. eller abstrakte begreber som

brugerne mener beskriver motivet. Der er i så fald tale om tags såsom "anticipation", "strength" og "excitement". Derudover er der en del tags som "vintage" og "old", der referer til fotografiernes og motivernes alder. Nogle af billederne appellerer mere til følelserne end andre. Eksempelvis er det henholdsvis ca. 19 og 21 % af tagsene tilføjet billede nr. 11 og 10, der tilhører denne kategori, mens det kun er ca. 4 % for billede nr. 23 og 25.

A5 (specifik person, dyr eller ting) er den fjerde bedst repræsenteret kategori. Her er det igen oplysninger fra posten, som mange af tagsene er baseret på. Eksempelvis er et billede med titlen "Jimmy Clabby. Boxing" (nr. 27) tagget med "jimmi", "clabby" og "Indiana wasp". Det er da også over halvdelen (ca. 56 %) af tagsene i denne kategori, der allerede er tilgængelige i posten. Derudover består mange af tagsene af registrerede varemærker, firmanavne og lignende, der er identificeret på billedet. Dette gælder for eksempel billede 16, der er tagget med "Arnsberger & Son", "Helmar Cigarettes", "Odol" osv. Dette er informationer, som ikke er tilgængelige via de nuværende poster.

Kategorierne A2 (generelt sted) og A3 (generel begivenhed eller handling) er nærmest ligeligt repræsenteret med henholdsvis ca. 6 og 5 % af det samlede antal tags i A-kategorien. Begge delkategorier består hovedsagligt af nye tags.

A4 (generel tid), A7 (specifik begivenhed eller handling), A8 (specifik tid), A10 (farve), der er de resterende i A-kategorien, er alle forholdsvis ringe repræsenteret og tæller således tilsammen kun for lidt over 8 % af A-kategorien og ca. 7 % af det samlede antal tags.

Størstedelen af de tags B-kategorien indeholder, er allerede registreret i posten. B3 (tekniske specifikationer) adskiller sig dog lidt, idet mange af nyhedsbillederne er tagget med "black and white" eller lignende, ligesom nogle af billederne er tagget med "large format". Begge dele er informationer, der ikke er tilgængelige via den nuværende post.

I C-kategorien er det C3 (sammensatte tags) og C4 (umulig at bestemme betydning), der er bedst repræsenteret. At der trods alt ikke er flere sammensatte tags end det er tilfældet, skyldes sandsynligvis, at det er muligt at separere ordene med mellemrum i Flickr. Det er derfor unødvendigt at lave sammensatte tags. C4 indeholder en blandet pose af ubestemmelige tags. Disse kunne sagtens være både selvrefererende eller organiserende, men medmindre man forhørte sig hos den pågældende tagger, er dette umuligt at vide. Det gælder for eksempel tags som "lee", "Gaul" og "2840", der er tilføjet henholdsvis billede nr. 14, 24

og 19. Derudover indeholder kategorien hjemmelavede ord eller slang såsom ”ginourmous” og ”oldtimey”.

C2 (organisation) er slet ikke repræsenteret og C1 (selvreference) kun ubetydeligt. Dette bekræfter antagelsen om, at brugerne ikke benytter denne type af tags, når de henholdsvis ikke er ejerne, og når der er tale om billeder.

Endeligt består næsten 6 % af det samlede antal tags af udenlandske ord (D2). Det er hovedsagligt (97 %) nyhedsbillederne, der er tagget med denne type. Det er svært at sige med sikkerhed, hvorfor der er så stor forskel på de to sæt billeder med hensyn til denne type tags. Det kan skyldes simpel tilfældighed. Der kan dog også argumenteres for, at nyhedsbillederne er af en mere international karakter, idet der ikke kun er billeder fra USA, men også eksempelvis Mexico, England og Libyen. Næsten en tredjedel af det samlede antal tags i denne kategori er tilføjet billede 25, der blandt andet er tagget med ordet ”Mexico” på over 10 forskellige sprog. Ellers er tagsene typisk på spansk og fransk.

Stavefejl (D1) og ”falske” ord (D3) udgør en lille del af det samlede antal tags. Af de 25 ”falske” ord, der er registreret i D3, tilhører over halvdelen endda billede nr. 18. Fotografiet forestiller bokseren Irving Melrose, som blev kendt under navnet ”Young Cy Young”. Han forveksles i dette tilfælde med den tidligere bokser ”Denton True ’Cy’ Young”, der i 1937 blev optaget i ”Baseball Hall of Fame” (Wikipedia bidragere, 2008b). Tags som ”baseball hall of fame” og ”boston red sox” er derfor forkerte, idet de ikke har noget med Irving Melrose at gøre. Derudover findes der baseballkort, der blev udgivet af tobaksvirksomheder, hvorfor dette billede er tagget med ”PIEDMONT CIGARETTES” og ”tobacco issue”. Billedet er dog ikke et eksempel på et sådan kort, hvorfor disse tags også er falske.

I det følgende afsnit diskuteres først resultaterne af analysen i forhold til både den relaterede litteratur og opgavens undersøgelsesspørgsmål. Dernæst diskuteres valget af metoden til undersøgelsen.

5. Diskussion

5.1 Diskussion af analysens resultater

Ifølge undersøgelserne (jf. afsnit 2.2) søger brugerne overvejende på specifikke geografiske steder (S3), hvilket svarer til A6 i denne undersøgelse. Selvom kategorien var andenbedst repræsenteret i denne undersøgelse, udgør tagsene kun ca. 13 % af det samlede antal tags

tilføjet billederne. Lidt under halvdelen af tagsene er dog nye, hvilket betyder, at posterne vil få næsten dobbelt så mange relevante emneindgange af den type som brugerne søger mest på, hvis tagsene fra Flickr blev implementeret. På den måde vil en implementering være meget værdiberigende.

Termerne er dog på forskellige specificitetsniveauer. I et eksempel som billede nr. 16 spænder niveauet fra "United States of America", som det mindst specifikke til "canarsie", der er et kvarter i Brooklyn og derfor en betegnelse nær det mest specifikke. Det kan diskuteres, hvorvidt dette er en fordel eller en ulempe, idet det blandt andet giver risiko for støj i søgningerne. I forhold til tags, der beskriver geografiske steder såsom disse, er det dog overvejende en fordel, da de bidrager med emneindgange, som kan tilgodese forskellige behov. Hvis en bruger eksempelvis ønskede bybilleder af New York og ikke kendte navnene på de forskellige bydele, ville det være et problem, hvis alle billederne var tildelt geografiske emneord på det mest specifikke niveau.

I Choi og Rasmussens (2003) undersøgelse var størstedelen af søgningerne på generelle handlinger (G2), hvilket svarer til A3 i denne undersøgelse. Umiddelbart udgør denne delkategori en meget lille del af det samlede antal tags. Størstedelen af tagsene er dog unikke, hvilket betyder, at de vil være værdiberigende for databasen.

Brugerne søger også på generelle ting, dyr og personer (G1), der svarer til A1 i denne undersøgelse. A1 er den delkategori, der indeholder størstedelen af tags. Derudover er langt hovedparten af tagsene nye emneindgange, der ikke er til stede i den nuværende post. Begge dele bidrager til, at en implementering vil være meget værdiberigende for LOC's database.

Mange af billederne er dog meget udtømmende tagget, så selvom der skabes flere emneindgange til de enkelte billeder, kan det også give risiko for støj i søgningerne. Billederne kan være så exhaustivt tildelt tags, at termerne ikke længere er beskrivende for motivet. Hvis en bruger søger efter et billede med termen "ring", forventer vedkomne at finde billeder, hvor en ring er en central del af motivet. Brugeren vil ikke forvente at enhver detalje af et billede er indekseret og derfor typisk søge på noget mere centralt i motivet. Det er dog ikke alle tagsene af denne type, der blot bidrager til støj i søgningen. Hvis en bruger eksempelvis søger et billede fra 1940'erne af et hus med "fishscale shingles", som er en form for tag, ville det med den nuværende database muligvis være nærmest umuligt.

I forhold til LOC's database vurderes det dog her som en fordel at der findes mange tags i denne kategori. Dette skyldes at posterne i LOC's database i mange tilfælde er tilføjet få emneord. Tagsene fra Flickr vil derfor i overvejende grad sikre en højere recall på trods af, at det i nogle tilfælde vil være på bekostning af en lav precision.

Kategorien indeholder også mange synonymmer. Dette har igen både fordele og ulemper, idet synonymmer kan give risiko for støj i søgningen og dermed forhindre et højt recall. På den anden side kan der dog argumenteres for, at de forskellige tags i højere grad modsvarer brugernes ordforråd end thesaurustermerne i LOC's database og derfor vil lette adgangen til billederne. Det er dog kun til en vis grad, idet det er tvivlsomt om brugerne af denne type af database ville søge på termer, der svarer til "strisser" eller "strømer". Dette hænger igen sammen med brugernes forventning om, hvilken slags termer billederne er indekseret med.

S1 (specifik person, dyr eller ting), som også er en type af termer, som brugergruppen overvejende søger på, svarer til A5 i denne undersøgelse. Selvom det ikke er så stor en procentdel af tagsene, som tilhører denne kategori, er lidt under halvdelen af tagsene dog nye emneindgange, som ikke er tilgængelige i de nuværende poster. Hvis man eksempelvis ønsker at se gamle reklameplakater som en del af bybilledet, vil de på nuværende tidspunkt være svære at finde, men ved en implementering af tagsene fra Flickr ville en sådan genfindingen være mulig. En tilføjelse af tagsene ville altså give flere både brugbare og relevante emneindgange til samlingen. Ud fra denne betragtning vil en implementering derfor være værdiberigende.

Brugerne af de sammenlignelige databaser søger til en vis grad også på specifikke datoer og perioder, hvilket svarer til A8 i denne undersøgelse. Denne kategori er ikke særlig godt repræsenteret af Flickr-tagsene og er ofte blot en gentagelse af informationerne i posten. Det er dog spørgsmålet om tagsene overhovedet ville kunne bidrage med mere information, når der eksempelvis allerede står "september" og "1941" registreret i posten. Behovet for specifikke tidsangivelser er derfor allerede opfyldt i den eksisterende post. Tagsene fra Flickr vil i dette tilfælde derfor ikke være værdiberigende.

En forholdsvis stor procentdel af tagsene fra Flickr tilhører about-kategorien (A9). Som undersøgelserne af brugernes søgeadfærd viste, benyttes denne type af termer meget sjældent af brugergruppen i deres søgeformuleringer. Det er dog en af de typiske typer i taggingsystemer (jf. afsnit 2.3), hvorfor den ikke overraskende er så godt repræsenteret her.

Selvom tagsene vil bidrage med flere både nye og brugbare emneindgange til basen, vil de i dette tilfælde ikke være værdiberigende. Dog kan der argumenteres for, at de vil hjælpe i en eventuel udvælgelsesfase, men det ligger udenfor rammerne af denne undersøgelse.

Resultaterne i B-kategorien var som antaget hovedsagligt en gentagelse af dataene i de eksisterende poster. Tagsene i denne kategori er derfor ikke værdiberigende for databasen, selvom de er brugbare.

Få af tagsene tilhørte C-kategorien. Dette er et udtryk for, at billederne ikke er 'egoistisk' tagget, som ellers Hammond og kollegaer (2005) argumenterede for at hoveddelen af billeder på Flickr er. Det skyldes formodentligt, at billederne ikke er brugernes egne. I dette tilfælde ville billederne sandsynligvis tilføjes som favorit, hvis ønsket var at lette en senere genfindning for den enkelte bruger. Der er dog få af tagsene, som muligvis kunne være et udtryk for selvreference (C1), men det er umuligt at vide, idet oplysninger om de forskellige taggere ikke umiddelbart er tilgængelig.

Der er ingen tags, der tilhører C2-kategorien, hvilket underbygger antagelsen om, at der ikke er så meget "task-organisation" med hensyn til billeder. Det er meget positivt i forhold til en eventuel implementering, at antallet af denne type tags ikke er højere, idet de ikke vil være hverken brugbare eller værdiberigende for databasen. Det er dog samtidigt lidt overraskende, idet denne type af tags er meget karakteristisk for taggingsystemer. I den lignende undersøgelse af Angus, Thelwall og Stuart (2008) var det alligevel 12 % af tagsene, der tilhørte denne type.

I samme undersøgelse (Angus, Thelwall & Stuart, 2008) var det også 12 % af tagsene, der var sammensatte, hvor det kun er knap 2 % i denne. Som før nævnt burde denne procentdel heller ikke være så stor, idet det er muligt at adskille de enkelte ord med et mellemrum på Flickr. Hvorfor forskellen på de to undersøgelser igen er så stor, er vanskelig at vurdere.

Mens 40 % af tagsene i Guy og Tonkins (2006) undersøgelse indeholdt stavefejl, var på et andet sprog end engelsk, sammensat eller bestod af koder, var det gældende for under 10 % af tagsene i nærværende undersøgelse. Forskellen kan skyldes undersøgelsesernes forskellige datagrundlag, hvor der i Guy og Tonkins (2006) undersøgelse var tale om en stikprøve af tags på hele Flickr. Hvis ejeren af billederne ikke er engelsksproget, er der stor sandsynlighed for at tagsene heller ikke er det. I denne undersøgelse, hvor der er tale om billeder fra det

amerikanske nationalbibliotek, er det naturligt at langt hovedparten af tagsene er på amerikansk/engelsk. At der er så få stavefejl (D1) og tags på andre sprog end engelsk (D2) betyder, at langt størstedelen af tagsene umiddelbart er brugbare. De få stavefejl kan desuden forholdsvis let rettes, så også disse tags vil kunne bruges i databasen.

De tags, som tilhører kategori D3 ("falske" tags), kan dog give problemer for validiteten af posterne, hvis de blev implementeret. Det er især ved billeder som nr. 18, hvor hovedparten af termene tilføjet er forkerte og hvor fejlen ikke er åbenlys. Et fotografi af et egern, som er tagget med "wedding", ville give støj i søgningen, men det ville være let at vurdere, at der ikke er tale om et bryllupsbillede. Dette har som konsekvens, at alle tagsene skal verificeres, hvis LOC vil være fuldkomne sikre på, at alle data tilføjet posterne er helt korrekte. Der er i så fald ikke længere tale om en helt så ressourcebesparende tilføjelse af metadata. Hvis der ses bort fra billede 18, var der dog meget få af de andre billeder, som var tagget med forkerte termer. Der kan derfor argumenteres for, at de enkelte "brodne kar" på et tidspunkt ville blive fundet af brugere, som så ville rapportere det, hvormed fejlene på denne måde løbende vil blive rettet. Spørgsmålet er om LOC kan acceptere risikoen for, at de søgbare data i posterne ikke er 100 % korrekte, når nu informationerne i bibliotekets kataloger helst skulle være korrekte og pålidelige.

5.2 Diskussion af metodevalget

Metoden til dataindsamlingen blev valgt for at sikre et bredt udsnit af billederne. Der kunne i stedet have været valgt at udtage en tilfældig stikprøve. Idet det er umuligt at sikre, at en stikprøve er 100 % repræsentativ, blev dette dog fravalgt. Metoden blev samtidig benyttet for at garantere, at der var et vist antal tags tilføjet hvert billede for at sikre et rimeligt analysegrundlag. Dette giver dog et lidt skævt billede af hvor mange tags, der er tilføjet billederne generelt. Selvom der til hovedparten af billederne er tilføjet rigtig mange tags, findes der dog billeder på profilen med under ti tags. Som følge af dataindsamlingsmetoden er de ikke repræsenteret i denne undersøgelse. Det vurderes dog, at dette ikke har en signifikant betydning for undersøgelsens resultater.

Selve kategoriseringsskemaet var overordnet set velvalgt i forhold til analysen af tagsene. Det kan dog diskuteres, hvorvidt kategorier såsom delkategori A1 skulle have været opdelt i nogle grader af specifikation, således at "fat woman" ikke kom i samme kasse som "woman".

Grunden til, at dette kan danne grundlag for en diskussion, er, at der både ligger en specifikation i førstnævnte term samt en subjektiv vurdering i forhold til sidstnævnte. Det ville dog gøre det sværere at sammenligne med brugernes søgeadfærd og at være objektiv i analyseprocessen.

Metoden generelt gør det ikke muligt at sige noget om, hvad brugernes motivation for tagging i denne kontekst er. Dette ville have været et meget interessant aspekt at inddrage, idet det ville give en klarere forståelse af typen af tags. Er motivet for tilføjelsen af tags eksempelvis at lette genfindingen for andre? Eller har brugerne blot tagget, fordi LOC har opfordret til det? Hvorfor vælger brugerne de tags, som de gør? Hvad er eksempelvis tanken bag at tilføje de mange synonymmer? Denne undersøgelse giver ikke mulighed for at svare på disse spørgsmål, men i forhold til problemformuleringen er det dog heller ikke nødvendige elementer.

6. Konklusion

80 % af de tags, som blev analyseret i denne undersøgelse, er et udtryk for billedernes emner. Størstedelen er tags, som beskriver generelle personer, dyr eller ting (A1). Dernæst beskriver hovedparten af tagsene specifikke steder (A6), billedernes aboutness (A9) og specifikke personer, dyr eller ting (A5).

I forhold til undersøgelserne af søgeformuleringerne i databaser tilsvarende LOC's, harmonerer typen af tags delvist. Brugerne søgte hovedsagligt på specifikke geografiske steder og dernæst på både generelle og specifikke ting, dyr og personer. Alle disse typer af emneord er godt repræsenteret af tagsene. Der er dog den store gruppe af tags, som beskriver billedernes aboutness, hvilket ikke er noget, som brugerne søger på.

Tagsene er i overvejende grad brugbare for databasen, idet der er meget få af den type tags, som ellers ofte benyttes i taggingsystemer. Her tænkes der specielt på de tags, som er et udtryk for selvreference og organisation eller som består af sammensatte ord. Derudover er der også få tags som indeholder stavfejl eller er på andre sprog end engelsk. Tagsene er derfor umiddelbar søgbare og kan benyttes til genfindning i databasen.

75 % af tagsene er termer, som ikke er til stede i de nuværende poster, hvilket alt andet lige betyder, at tagsene også vil være værdiberigende ved en implementering.

Det, som taler imod en implementering er, at billederne er tildelt tags så exhaustivt, at det giver risiko for støj i søgningerne. Derudover er der mange synonymer og termer på forskellige specificitetsniveauer tilføjet. Endeligt er en del af tagsene ”falske”, hvilket giver problemer for validiteten af posterne.

I opgaven argumenteres det for at fordelene ved en implementering generelt opvejer ulemperne, idet tagsene bidrager med emneindgange af en type, som brugerne søger på, at synonymerne kan afspejle brugernes ordforråd i højere grad end thesaurustermerne og at flere specificitetsniveauer kan tilgodese forskellig brug. Med hensyn til de falske tags, er det dog et ægte problem, der kun kan løses ved at samtlige tags verificeres før en implementering. Dette vil dog betyde at implementeringen vil blive knap så ressourcebesparende og det er i så fald et spørgsmål om den kan betale sig. Da det er et fåtal af billederne, som er indekseret med denne type af tags, argumenteres der i opgaven for at brugerne med tiden vil finde fejlene og de på den måde vil blive rettet løbende af LOC. Det er dog ikke en optimal løsning.

Den endelige konklusion på undersøgelsen er derfor at en umiddelbar implementering af tagsene fra Flickr i Library of Congress' billeddatabase vil være fordelagtig, hvis LOC kan acceptere risikoen for, at få af billederne indeholder fejlagtige informationer.

7. Perspektivering

Resultaterne fra undersøgelsen viser, at selvom tags ikke følger de traditionelle regler for indeksering, kan de i nogle tilfælde med fordel benyttes som emneord alligevel. Det kunne være interessant at undersøge om kvalitet af tagsene til dette formål ville blive forhøjet, hvis man benyttede tagging-vejledninger (jv. afsnit 2.3). Ved eksempelvis at bede brugerne tage billederne med de termer, som de synes beskriver billedet bedst eller beskriver noget signifikant ved billedet, kan man muligvis undgå den høje exhaustivitet. Hvis man derudover gør opmærksom på, hvad formålet med tilføjelsen af tagsene er, vil man muligvis også sikre, at tagsene bliver endnu mere anvendelige. Endeligt kunne det være en fordel at benytte et taggingsystem, der gav mulighed for at tage den samme ressource med det samme tag flere gange, for på den måde at kunne se hvilke tags flest brugere tilføjer.

8. Litteraturliste

- Andersen, J. *et. al.* (2006). Indeksering. I: *Informationsordbogen : ordbog for informationshåndtering, bog og bibliotek*. Lokaliseret 20.05.08 på <http://informationsordbogen.dk/concept.php?cid=876>
- Angus, E., Thelwall, M. & Stuart, D. (2008). General patterns of tag usage among university groups in Flickr. *Online Information Review*, 32(1), 89-101.
- Armitage, L. & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science*, 23(4), 287-299.
- Becker-Christensen, C. (1999). Brugbar. I: *Politikens Nudansk Ordbog* (17. udg, 1. opl.). København: Politikens Forlag A/S. (dansk-dansk ordbøger).
- Choi, Y. & Rasmussen, E. (2003). Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. *Journal of the American Society for Information Science and Technology*, 54(6), 498-511.
- Collins, K. (1998). Providing Subject Access to Images: A Study of User Queries. *The American Archivist*, 61, 36-55
- Cox, A. M., Clough, P. D. & Marlow, J. (2008). Flickr : A first look at user behaviour in the context of photography as serious leisure. *Information Research*, 13(1), paper 336. Lokaliseret 08.04.08 på <http://InformationR.net/ir/13-1/paper336.html>.
- Enser, P. G. B. (1993). Query analysis in a visual information retrieval context. *Journal of document & text management*, 1(1), 25-52.
- Enser, P. G. B. & McGregor, C. G. (1993). *Analysis of visual information retrieval queries* (6104). London: British Library.
- Enser P. G. B. & Sandom, C. J. (2007). Facing the reality of semantic image retrieval. *Journal of Documentation*, 63(4), 465-481.
- Farkas, M. G. (2007). What is social software? I: M. G. Farkas, *Social software in libraries : building collaboration, communication, and community online*. New Jersey: Information Today, Inc. (s. 1-9).
- Frost, C. O. *et. al.* (2000). Browse and Search Patterns in a Digital Image Database. *Information Retrieval*, 1, 287-313.
- Golder, S. & Huberman, B. A. (2005). The structure of Collaborative Tagging Systems. *Technical report, In-formation Dynamics Lab, HP Labs*. Lokaliseret 21.04.08 på <http://arxiv.org/ftp/cs/papers/0508/0508082.pdf>.
- Guy, M. & Tonkin, E. (2006). Folksonomies : Tidying up tags? *D-Lib Magazine*, 12(1). Lokaliseret 21.04.08 på <http://www.dlib.org/dlib/january06/guy/01guy.html>.
- Hammond, T. *et. al.* (2005). Social bookmarking tools (I) : a general overview. *D-Lib Magazine*, 11(4). Lokaliseret 21.04.08 på <http://www.dlib.org/dlib/april05/hammond/04hammond.html>.
- Jørgensen, C. (1999). Access to Pictorial Material : A Review of Current Research and Future Prospects. *Computers and the Humanities*, 33, 293-318.

- Kipp, M. & Campbell, D. (2006). Patterns and inconsistencies in collaborative tagging systems : an examination of tagging practices. I: *Proceedings Annual General Meeting of the American Society for Information Science and Technology*, Austin, Texas (US). Lokaliseret 21.04.08 på <http://eprints.rclis.org/archive/00008315/fullmetadata.html>.
- Library of Congress (1995). *Thesaurus for graphic materials* (2. rev. udg.). Washington, D.C.: Cataloging Distribution Service, Library of Congress. Heri: afsnit I.A og III.
- Library of Congress (2007). *About the Prints & Photographs Online Catalog*. Lokaliseret 19.05.08 på <http://www.loc.gov/rr/print/catalogabt.html#scope>.
- Macgregor, G. & McCulloch, E. (2006). Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool. *Library Review*, 55(5), 291-300.
- Marlow, C. et al. (2006). Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. *Proceedings of the WWW 2006 Collaborative Web Tagging Workshop, 2006*.
- Mathes, A. (2005). Folksonomies : Cooperative Classification and Communication Through Shared Metadata. *Computer Mediated Communication – LIS590CMC, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December 2004*.
- Matusiak, K. K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services : International digital library perspectives*, 22(4), 283-298.
- Panofsky, E. (1972). *Studies in iconology: humanistic themes in the art of the Renaissance*. Oxford, UK: Westview Press.
- Rafferty, P. & Hilderly, R. (2007). Flickr and Democratic Indexing : dialogic approaches to indexing. *Aslib Proceedings: New Information Perspectives*, 59(4/5), 397-410.
- Ranganathan, S. R. (1962). *Elements of Library Classification* (3. ed.). Bombay, India: Asia Publishing.
- Rasmussen, E. M. (1997). Indexing Images. *Annual Review of Information Science and Technology (ARIST)*, 32, 169-196.
- Raymond, M. (2008a). *My Friend Flickr: A Match Made in Photo Heaven*. Library of Congress Blog. Lokaliseret 08.04.08 på <http://www.loc.gov/blog/?p=233>.
- Raymond, M. (2008b). *More Photos in Flickr*. Library of Congress Blog. Lokaliseret 08.04.08 på <http://www.loc.gov/blog/?p=268>.
- Shatford, S. (1986). Analyzing the Subject of a Picture : A Theoretical Approach. *Cataloging & Classification Quarterly*, 6(3), 39-62.
- Shatford Layne, S. (1994). Some Issues in the Indexing of Images. *Journal of the American Society for Information Science*, 45(8), 583-588.
- Shirky, C. (2005). *Ontology is overrated : categories, links and tags*. Clay Shirky's Writings About the Internet, Shirky.com. Lokaliseret 22.04.08 på http://shirky.com/writings/ontology_overrated.html.
- Svenonius, E. (1994). Access to Nonbook Materials : The Limits of Subject Indexing for Visual and Aural Languages. *Journal of the American Society for Information Science*, 45(8), 600-606.

Tennis, J. T. (2006). Social tagging and the next steps for indexing. *17th SIG/CR Classification Research Workshop, 4. November 2006*.

van Hooland, S. (2006). From Spectator to Annotator : Possibilities offered by User-Generated Metadata for Digital Cultural Heritage Collections. *CILIP Cataloguing Indexing Group Annual Conference, University of East Anglia, 13.-15. september 2006*.

Voss, J. (2007). Tagging, Folksonomy & Co : Renaissance of Manuel Indexing? I: *the 10th international Symposium for Information Science, Cologne, Germany*. Lokaliseret 26.04.08 på <http://arxiv.org/abs/cs/0701072v2>.

Wikipedia bidragere (2008a). Flickr. I: *Wikipedia : The Free Encyclopedia*. Lokaliseret 16.05.08 på <http://en.wikipedia.org/w/index.php?title=Flickr&oldid=212460183>.

Wikipedia bidragere (2008b). Cy Young. I: *Wikipedia : The Free Encyclopedia*. Lokaliseret 20.05.08 på http://en.wikipedia.org/w/index.php?title=Cy_Young&oldid=213530100

Websider:

Del.icio.us: <http://del.icio.us>

Facebook: <http://www.facebook.com>

Flickr: <http://www.flickr.com>

Library of Congress' profil på Flickr: http://www.flickr.com/people/library_of_congress/

MySpace: <http://www.myspace.com>

Prints & Photographs Online Catalog: <http://lcweb2.loc.gov/pp/pphome.html>