Decentralized subject indexing of television programs: the effects of using a semi-controlled indexing language

Veslemøy Søbak and Nils Pharo*

Oslo and Akershus University College
of Applied Sciences
PO box 4 St. Olavs plass
NO-0130 Oslo, Norway

veslemoy.sobak@gmail.com; nils.pharo@hioa.no

tel: +47 22452684; fax: +47 22452605

*corresponding author

Abstract

We have performed an exploratory case study to understand how subject indexing performed by television production staff using a semi-controlled vocabulary affects indexing quality. In the study, we used triangulation, combining tag analysis and semi-structured interviews with production staff of the Norwegian Broadcasting Corporation. The main findings reveal incomplete indexing of TV-programs and their parts, in addition to low indexing consistency and uneven indexing exhaustivity. The informants expressed low motivation and a high level of uncertainty regarding the task. Internal guidelines and high domain knowledge among the indexers does not form a sufficient basis for creating quality and consistency in the vocabulary. The challenges that are revealed in the terminological analysis, combined with low indexing knowledge and lack of motivation, will create difficulties in the retrieval phase.

Introduction

Both internal and external access to television programs depends on indexing. Currently many broadcasters make their programs digitally available for the general audience via the web, but in order to retrieve a particular news story or a part of a talk show or a song played in a live concert, we need someone to describe where to find it. It is of course also essential for the program creators, archivists and other internal parties to keep track of individual programs and program parts, for example to reuse the content in other programs.

Two different approaches to image retrieval, both of moving and still images, exist; text based and content based. Text based image retrieval is based on textual descriptions of the images whereas in content based image retrieval characteristics such as color, shape and texture are extracted from the images and matched with queries in the same format (Goodrum, 2000). In the current study, we examine text-based indexing of TV programs.

High quality indexing forms the basis for effective information retrieval. Traditionally, librarians and other information managers have performed subject indexing of books and other material. With the advent of web technology, however, we have seen the development of a myriad of information services where indexing also is performed by the users of the services in the form of tagging. In our

study, we have investigated the situation where indexing has been *decentralized* from the information specialists to be performed by persons involved in the production process, which constitute a third category of indexers. This group of indexers have high domain knowledge, but are not professionally trained indexers. Content creators as indexers is known from scientific publishing, where authors are commonly asked to provide keywords describing their texts, but as far as we know this is new within television broadcasting.

If we look upon the selection of index terms on a scale from "fully controlled vocabularies" to "folksonomies" (Wal, 2007) we will often find information services indexed by librarians and end users on each end of the scale. We wanted to investigate indexing by resource creators who had access to all previously used indexing terms, and who were allowed to add the terms they found most appropriate to describe the content. New terms were thereafter controlled by indexing experts, who either approved them or added alternative terms. We have coined the term *semi-controlled vocabulary* to describe this type of indexing vocabulary.

The indexers need to follow an indexing policy developed to secure effective retrieval. On our indexing scale they will thus be somewhere between the two extremes, and it is our belief that the findings will provide valuable information for development of indexing policies and practice. Our case study was performed in the Norwegian Broadcasting Corporation (NRK), where we have examined the indexing of television programs over a one-week period.

We formulated the following general research question: *What consequences does decentralized indexing have on the indexing vocabulary?*

We specifically want to address some consequences operationalized in the form of the following specific research questions:

RQ1: What characteristics of television programs are indexed?

RQ2: How does decentralized indexing influence indexing consistency?

RQ3: How do indexers' practice and motivation affect indexing quality?

In the next section we examine previous work on indexing, thereafter follows our methods, before we present our findings. The final section contains conclusions and discussion.

Previous work

Subject indexing is a two-step process, a conceptual content analysis is followed by a translation to the appropriate terms in a chosen vocabulary (Lancaster, 2003). An alternative approach to alphabetical subject languages is to use a classification scheme, such as the Dewey Decimal Classification System. We will not deal with the latter here.

Different types of controlled vocabularies for subject indexing exist, varying with respect to their "semantic strength", i.e. their ability to express terminological relationships. The simplest subject indexing lists would contain no relationships, but most lists of "subject headings", e.g. Library of Congress Subject Headings, have a hierarchical structure and are thus taxonomies (Broughton, 2004). Thesauri in addition contain associate relationships as well as synonym control in the form of equivalence relationships. In contrast to these we find uncontrolled indexing, of which user created tags have received a lot of attention in indexing research the last 10 years. An important goal of indexing with controlled vocabularies is to improve consistency in indexing. This means that even if the same topic is written about by several authors using different terms indexing with a controlled

vocabulary will ensure that all these documents will be retrieved when needed. If uncontrolled terms are used consistency will drop

From the literature we know that indexing quality foremost depends on the consistency of indexing, i.e. the consistency in interpreting the content as well as the consistency in choosing the appropriate terms. In addition indexing exhaustivity, i.e. to what degree all topical aspects of the document's content is indexed, often referred to as "breadth" of indexing, affects indexing consistency, and hence quality. The more terms that are used to index a document the more probably consistency will suffer (Lancaster, 2003). The specificity of the indexing language also affects consistency. As specificity of terms increase the number of terms in the language will also increase, which in turn will make it more difficult for indexers to use terms consistently (Cleverdon & Keen, 1966; Jones, 1972).

If no controlled vocabulary is used there is a risk of "messy metadata" (Smith, 2008), i.e. several versions of the same term are used, e.g. in singular and plural form, and synonyms and near-synonyms are used to describe the same topic. Peters (2009) summarized the most common advantages (and disadvantages) of tagging/folksonomies. The 'vocabulary problem' denotes the problems users, in particular novices, have in finding the correct terms to express their needs. One possible advantage of a folksonomy is that it "authentically reflect the users' language and thus solve the 'vocabulary problem'" (Peters, 2009, p. 218). Further, a folksonomy allows for different interpretation and hence allows more terms for representing the same phenomenon, provides more access points to the information resources and makes it easier to index new phenomena.

Not much literature can be found on manual subject indexing of television programs, but, depending on the exhaustivity of indexing, it can be a very time-consuming task. According to (Dowman, Tablan, Cunningham and Popov (2005, p. 225) "it takes a BBC archivist almost seven hours to catalog Newsnight, a fifty minute daily news broadcast, in detail". In an experimental study Laine-Hernandez and Westman (2006) analyzed indexers (technology students and university staff) that assigned uncontrolled keywords to, and wrote, free descriptions of journalistic images. They found that 26.3 % of the keywords described "objects" whereas 28.2 % described the content or "story" attributed to the images. Other keyword classes included "people-related attributes" (12.1 %), "abstract concepts" (10.8 %) and "visual elements" (7.2 %). Thereafter the participants categorized the images. Frequently occurring categories were religion, animals, politics, scenery, sports and music. Markkula and Sormunen (2000) studies the difference in news photos indexed on various levels by professionals. Photos were partly indexed using a 'free description field' (for describing concrete objects) partly with a thesaurus (for 'conceptual indexing'). The researchers observed that indexing was quite inconsistent, one problem being that "that controlled indexing was applied only to a share of photos" (Markkula & Sormunen, 2000, p. 20). A study of journalists searching the photo collection revealed unsatisfactory results, in particular when their information needs went beyond looking for photos of persons or other named objects. A study by Turner (1995) reports a high degree of consistency in user and indexer-assigned terms to a selection of moving images from the National Film Board of Canada's stock shot collection. A project at the University of Texas (Geisler, Willard, & Ovalle, 2011) suggests a crowdsourcing framework for indexing film and television, but the framework seems to be targeted at works of fiction rather than e.g. news programs. A few studies of how users tag YouTube videos have been reported, Knautz and Stock (2011) found that users in general consistently indexes emotions in video. Knautz and Stock's study only involved nine classes of emotions and no other dimensions of the videos were indexed. Agius, Angelides and Zad (2012) compared different tools used for indexing multimedia resources, more specifically unstructured folksonomy (Flickr, YouTube and del.icio.us) and a structured Mpeg7 tool. They found that using a structured tool "ensures that the tagging process results in very comprehensive and clear tags"(Agius

et al., 2012, p. 167), however, this is at the expense of a considerably more complicated tagging process.

A substantial portion of work has, on the other hand, been done on content based indexing of moving images (Lew, Sebe, Djeraba, & Jain, 2006). An overview of the technical aspects of video retrieval can be found in (Smeaton, 2004). Automatic speech recognition is another technique tested out for indexing television and radio programs (Dowman et al., 2005). Automatic genre classification of TV programs (Montagnuolo & Messina, 2007) and YouTube videos (Ekenel & Semela, 2011) based on their visual and audio-based cues also exist.

Since news constitute a large share of television content it is relevant to take into account news indexing, several vocabularies have been developed specifically for describing news items, such as the IPCT NewsCodes[1] and PBS Merlin Topic Taxonomy[2]. Three major German and Austrian television channels have published their guidelines for indexing (ARD/ORF/ZDF, 2008) the content of programs. Research is, however, mainly focused on automatic classification and text mining of news items (e.g. Rocha & Cobo, 2011) rather than conceptual indexing. Indeed, news material from early on constituted an important part of IR test collections, such as the TREC collections, the first example being Salton's Time magazine collection (Sanderson, 2010).

To supplement our understanding of subject indexing performed by non-information professionals we have examined studies on the purpose of tagging, studies of tagging and indexing consistency and studies comparing different types of indexers.

The purpose of tagging

Golder and Huberman (2006) have studied tagging and usage patterns in the bookmark-sharing site Delicious and found that the tags represent different functions for the users. They claim that "[t]agging is an act of organizing through labeling, a way of making sense of many discrete, varied items according to their meaning" (Golder & Huberman, 2006, p. 203) and categorize tags into seven categories according to topicality, type of item (e.g. blog), ownership, tags refining categories, quality or characteristics (e.g. scary), self reference (e.g. mystuff) and organization (e.g. toread). Topical tags constitute the largest group of tags and the authors distinguishes between the first four categories, classifying them as non-personal or tagger extrinsic and the latter three, which are "only relevant to the tagger" (Golder & Huberman, 2006, p. 204). A study performed by by Marlow, Naaman, Boyd and Davis (2006), reveals a complementary picture. They have created a taxonomy of user incentives for tagging. The taxonomy contains six somewhat overlapping categories of incentives: future retrieval; contribution and sharing; attract attention; play and competition; self presentation and opinion expression. Thus topical tags can be the result of different intentions from the tagger, and these intentions might differ over time even if the resulting tags are the same. Kipp (2007) did a study of 78 personal tags, i.e. tags that did not describe the item's topic, in the three bookmarking sites CiteULike, Connotea and Delicious. She categorized the tags into two broad categories: "Time, task or project related tags" (48 tags) and "Affective tags" (30 tags). The majority of tags in the first category are variations of the "ToRead" tag, i.e. tags used for personal document management. Typical examples of affective tags are "cool" and "fun", representing Marlow et al.'s "opinion expression" category and Golder and Huberman's category "quality or characteristics". In a study where users of Flickr and Zonetag (a Flickr mobile application) were interviewed Ames and Naaman (2007) identified four different motivational factors. Self/organization represented tagging performed by users who

[1] http://cv.iptc.org/newscodes/subjectcode/
[2] https://projects.pbs.org/confluence/display/merlin/Topic+Taxonomy

focus on images they had retrieved; self/communication, tagging performed to represent memories and context; social/organization, where users tag to make it possible for others to find their images; and social/communication, which represents the intention of communication and pointing out the context of the pictures for others. The authors find that tagging for the public is the main motivation for most users. "Family and friends" is added as a fifth target category in (Nov, Naaman, & Ye, 2008), who, using surveys, find that there is a positive correlation between level of users "self" and "public" motivation and the number of tags. They are not able to find any correlation between "family and friends" and number of tags. Sa and Yuan (2013) in a small study of users of five different tagging sites (last.fm, flickr, delicious, citeulike and movielens) found that the most important motivation for creating tags was to facilitate future retrieval (by themselves and others), for using tags the most important motivation was, not unsurprisingly, to "search".

Tagging and indexing consistency

Kipp and Campbell (2006) analyzed tagging patterns in order to investigate how user-created tags support traditional indexing methods. 165 831 Delicious tags were the subject of co-word analysis and the authors found that tag relationships "do not always follow the co-word clusters, [but] … often follow relationships of synonymy" (Kipp & Campbell, 2006, p. 15). Interestingly enough since solving synonymy problems is perhaps the most important objective of controlled vocabularies. Spiteri (2007) performed a linguistic analysis of user-generated tags in Delicious, Furl and Technorati, which she compared to NISO's guidelines for controlled vocabularies. She found that the tags in general conform to the guidelines with respect to their grammatical form (e.g. 94 to 97 % of all tags were nouns or noun phrases), but that inconsistencies existed in the use of singular and plural form. She considers the most problematic area to be the use of ambiguous terms such as homographs, abbreviations and acronyms. Guy and Tonkin (2006) has performed a spelling control of Delicious and Flickr tags and found that 28 % of Delicious tags and 40 % of Flickr tags did not match the terms of a grammatical database, due to spelling errors, but also because the terms were compound words consisting of more than two terms, contained numbers or were in plural form.

Comparison of indexers

Margaret Kipp has performed several studies of the indexing practice of three types of indexers: users, authors and professional indexers. In (Kipp, 2005) she studied similarities and differences in the terms used for indexing academic articles by the three indexer groups. She found that the terms used for a majority of the articles (133 out of 165) are related, but not according to the guidelines formally prescribed for thesauri. There are also conceptual differences between the terms set by the different indexer groups; the terms used by the users are of a more general nature than professionals' tags. Two studies published in 2011, one within library and information science (Kipp, 2011b), the other in medicine (Kipp, 2011a) verify that there are clear differences in the different groups' tagging behavior, stating that "[w]hile tags and author keywords were found that matched descriptors exactly, other terms which did not match but provided important expansion to the indexing" (Kipp, 2011a, p. 245). Similar results have been reported for indexing of books (Thomas, Caudle, & Schmitz, 2009). Recently, Bogers and Petras (2015) performed a study comparing which method is more helpful for book searching, tagging or controlled vocabularies? Their conclusion was that the two methods complement each other, e.g. in that tags seem to perform better for retrieving fiction than terms from controlled vocabularies. In addition, they find that more unique controlled vocabulary terms are used for indexing, whereas tags are more often repeated. Current research thus suggests that the use of more unique terms, independent of these coming from a controlled vocabulary or not, are good for improving precision and controlled vocabulary terms improve recall during book retrieval.

Method

We have performed an exploratory case study of subject indexing performed in the Norwegian Broadcasting Corporation (NRK). NRK is the largest media organization in Norway, founded in 1933 with approximately 3500 employees. It is a publicly funded, public-server broadcaster, running three national television channels and three national radio channels in addition to several digital radio stations. In 2007, the organization introduced a new tool for managing programs. All metadata of radio and television broadcasts, including subject terms (hereafter called "tags"), are registered in the program bank. It was also decided that metadata was to be registered by staff involved with program production. Until then professional media archivists had handled this. The goal is to establish a metadata regime that couples content-wise similar material across different media formats. With the advent of a new web based media player, it was also the intent to include user based tagging of the programs. From 2012, the production staff supervised by the archive and research (A&R) department have performed all metadata registration. NRK is the first public broadcaster to implement such a regime. User based tagging has not yet been implemented. The goal of the new metadata and indexing policy is to make this content available for end users at the same time the content must be managed for internal reuse. For each program, seven metadata fields are recorded, either automatically or manually: title; date of broadcast; host/team; participants; introduction/heading; tags; and rights. The tags are not from a controlled vocabulary, but the indexer has access to previously used tags, and similar terms from the list are automatically suggested during indexing. This vocabulary consists of approximately 7000 terms created by the indexers. After a program is indexed the metadata is controlled by staff in the A&R department. If a new tag is accepted, it is added to the vocabulary, alternatively A&R selects a term from the vocabulary.

The A&R department has developed an indexing policy in the form of tagging guidelines, stating that:

1. Tags should cover the "who", "what" and "where" of all indexed program parts. "When" is included when necessary
2. Tags are written in bokmål (one of the two standard forms of the Norwegian language)
3. Use single terms and concepts should be used, not sentences
4. Use common terms and concepts
5. Well-known names on incidents should be used
6. Abbreviation are used if these are better known
7. Common synonyms should be added
8. Terms must be precise, but more general terms can be added as well
9. Ambiguous terms should not be used
10. Tags should always be chosen with caution[3]

In addition the guidelines states that lower case letters should be used and advice indexers to pay attention to the use of singular and plural form of substantives, pointing out particular cases when one or both forms should be used[4].

Tags are added both to the program as a whole and to different program parts (such as the individual news items in the daily news program). Compared to the previous indexing policy, centralized and done by trained indexers, the biggest difference is that tags replaced Dewey classification codes for image description and content description.

[3] For example: the tag "murderer" should not be used before the accused person is convicted

[4] For example: if a substantive has irregular plural form both forms are used

In order to answer our research questions we have performed an analysis of a selection of tags and performed interviews with metadata registering production staff.

Tag analysis

We collected all tags on program parts, broadcast by NRK1 (the broadcaster's main channel) during one week in October 2012. The metadata database did not include tags added to programs as a whole. During this week, the channel showed 249 programs with a total of 823 program parts. Program IDs were looked up in an internal database and used to retrieve all metadata, including the tags. We collected data on broadcasting date and time, program ID, title, introduction/heading, editorial staff category (e.g. news or sports) and tags into a spreadsheet. All program reruns (within the same week) as well as programs created by NRK's regional offices were excluded because they were not available in the metadata database. Program parts were identified, since these sometimes are reused, typically in news and sports programs, and given IDs and titles.

The tags were manually categorized by one of the researchers, following an inductive procedure. Additional metadata fields, such as titles and introductions, were used to help decide the context of tags. The researcher developed a facetted schema of tag categories during the categorization. This consists of seven main categories (see Figure 1). The categories are mutually exclusive. The categories emerged from the data and tags were later reanalyzed by the same researcher to identify possible subcategories. This way we were also able to check the consistency of the first categorization round. The re-analysis did not result in any tags being put into a new main category. In all 23 subcategories were created, ten of which were divisions under topic and six under place.

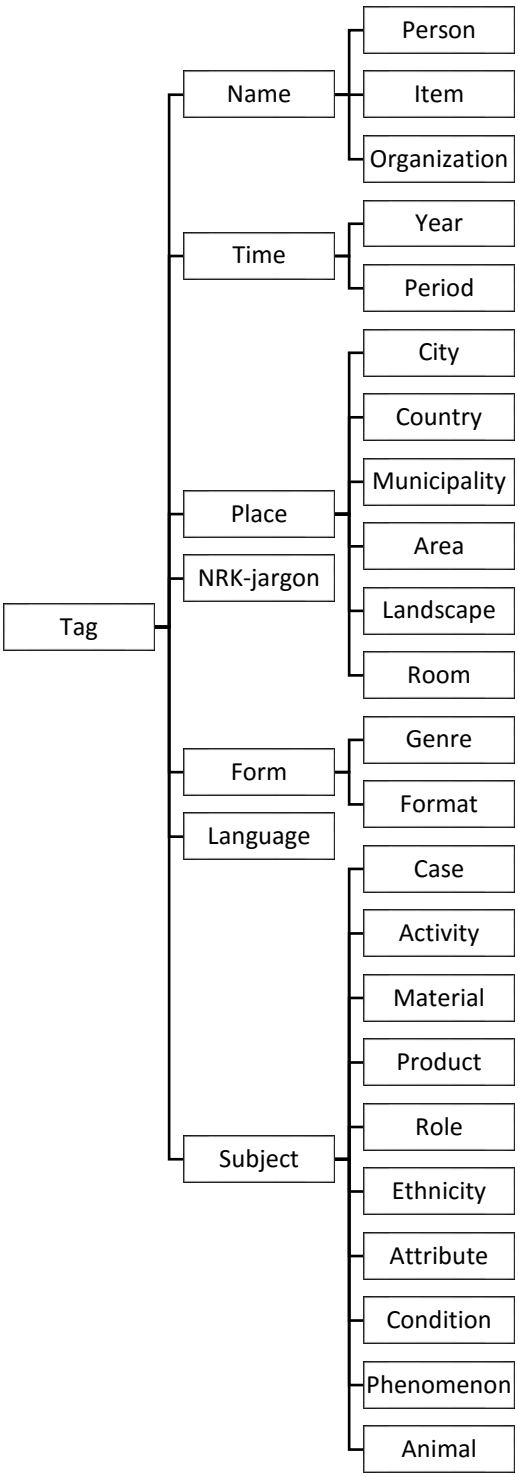

Figure 1 Categories emerging from the tag analysis

Interviews

In order to understand the motivation and attitudes towards tagging among the production staff we conducted informal interviews with members of different editorial offices. The interviews were scheduled to take place at the interviewees' workplace at a time when they were involved with metadata registering. The interviews started with each informant showing how they indexed a program, and explaining their routines. We also asked the informant to show us an example of their tagging procedure. The latter for us to become more familiar with the process.

We developed an interview guide with the intent of having the informants tell how they performed their tagging, how they chose the appropriate terms, the level of description, target groups and focus, and if they followed particular procedures or patterns when tagging. Thereafter they were asked questions about their experiences with and attitudes towards tagging, e.g. how they assessed their own tagging competence, their understanding of the purpose and function of tags, their training and attitudes towards the tagging guidelines, and their own experience with the tagging tools and the online media player.

We wanted to interview five members of the production staff representing different editorial offices and with different roles in the production. In all, we contacted 41 staff members to get the necessary number of interviewees. All interviews took place during one week in February 2013. All interviews followed the guide, but some modifications were made to the guide after the first interview. These included changing the order of the questions and adding a question about the involvement with the program production (this information came voluntarily from the first interviewee). The interviews lasted from half an hour to one and a half hour.

Each interview was recorded and transcribed immediately after the conversation; in addition, we made a short abstract based on notes made by the interviewer. We used "Meaning condensation" (Kvale, 2007) when this was appropriate to compress the informants' utterances by removing irrelevant information and extract the essence in what is said. Shorter and concise answers were kept in their original form.

Findings

We will now answer our specific research questions, and we start with the tag analysis in order to find out what aspects of TV programs are indexed. Next, we examine indexing consistency and the terms used in the uncontrolled vocabulary. Lastly, we inspect the indexing behavior and motivation of the indexers and analyze how this affects the indexing.

RQ1: What characteristics of television programs are indexed?

From October 8th to October 14th 249 programs were broadcast by NRK, of these 9 programs made by the regional offices and 49 duplicates (reruns sent during the same period) were removed. 135 programs did not contain tags, thus we ended up with 56 programs to analyze, i.e. 22 % of all programs sent in the period (see Table 1).

Table 1 Tags distributed per day on program level

Overview	**8 oct**	**9 oct**	**10 oct**	**11 oct**	**12 oct**	**13 oct**	**14 oct**	**Total**
No of programs	33	43	39	38	35	31	30	249
Reruns	2	5	5	5	3	17	12	49
Regional news	2	2	2	2	1	0	0	9
Programs w/o tags	21	23	25	22	21	10	13	135
Analyzed programs	8	13	7	9	10	4	5	56

The 56 programs contained 823 program parts, of which 447 did not contain any tags. In addition five parts were duplicates, i.e. the same part was registered one or more times with identical metadata (including tags). This leaves us with 371 tagged program parts with a total of 1828 tags (see Table 2).

Table 2 Tags distributed per day on program part level

Overview	8 oct	9 oct	10 oct	11 oct	12 oct	13 oct	14 oct	Total
No of program parts	124	193	145	143	147	35	36	823
Duplicates	0	4	1	0	0	0	0	5
No of parts w/o tags	61	91	83	85	94	11	22	447
Analyzed program parts	63	98	61	58	53	24	14	371
No of tags	277	488	284	314	263	126	76	1828

The tags were distributed on program parts created by six different editorial offices (Table 3). It is important to be aware that the analyzed channel broadcasts very few programs targeting children, since 2007 NRK has a dedicated children's channel. This explains the low number of tags from this editorial office.

Table 3 Tags distribution on editorial offices

Editorial office	No of tags	Percentage
News	758	41.5 %
Culture and entertainment	561	30.7 %
Sport	283	15.5 %
Documentaries and fact	179	9.8 %
Health, consumer and life style	31	1.7 %
Children	16	0.9 %
Total	1828	100 %

We have categorized the tags according to the category scheme in Figure 1. 61.5 % (1125 out of 1828) of all tags were *Subject* tags, approximately 26 % represented *Names*, and 8 % were categorized as *Place* tags. The other categories (Form, Time, Language and NRK-jargon) each represented approximately 1 % of all tags. In addition, 18 tags were impossible to categorize because of ambiguity, i.e. "suggestion" (forslag) and "more" (mere). Places were divided into geographical locations, such as specific countries and cities and "places of residence", which included nursing home, shoreline and crossroads. The Name category included personal names (including the name of fictive persons such as "Han Solo"), names of organizations as well as names of items (e.g. song titles).

In Table 4, we present the subcategories emerging from our material. Product and Activity are the two most common subject tags, whereas Organization and Person dominate the name category. Examples of Activity tags are "homework help" (leksehjelp), "espionage" (spionasje) and "boiling eggs" (koke egg). The *Product* subcategory includes idiosyncratic tags like "robot window cleaner" (robotvindusvasker) and "veteran motor bike" (veteranmotorsykkel), as well as "telephones" (telefoner), "dinner" (middag) and "budget" (budsjett). 152 tags are categorized as *Role*, including "polar heroine" (polarheltinne), "meat wholesalers" (kjøttgrossister) and "preliminary minister" (settestatsråd). Other categories with a significant number of tags include *Condition* ("underdog"), *Animal* ("tadpole") and *Case*. The latter was used for terms representing very general topics, such as "soccer" and "technology". The remaining subject tags are relatively seldom used, but do represent distinct tags that did not belong to any of the other categories. Named organizations include the football team "Real Madrid" whereas the person "Lance Armstrong" was tagged five times in the period. *Activity* is used to categorize tags were an activity, such as "frying" (steiking) or "filming", is

treated as a subject in the TV program, it is not a characteristic of the program as such (which would be comparable to Ranganathan's 'energy' category from his PMEST schema). Among the NRK-jargon tags, we find one example of such a tag, the photography technique "panning" (panorering).

Table 4 Distribution of tag in categories and subcategories

Category	**Subcategory**	**No of tags**	**Percentage**
Subject	Product	389	21.3
Subject	Activity	340	18.6
Subject	Role	152	8.3
Subject	Case	69	3.8
Subject	Condition	53	2.9
Subject	Animal	45	2.5
Subject	Attribute	32	1.8
Subject	Phenomenon	19	1.0
Subject	Ethnicity	21	1.1
Subject	Material	5	0.3
Place	Institution	35	1.9
Place	Area	27	1.5
Place	Country	27	1.5
Place	Landscape	19	1.0
Place	City	16	0.9
Place	Room	12	0.7
Place	Municipality	11	0.6
Name	Organization	214	11.7
Name	Person	178	9.7
Time	Period	10	0.5
Time	Year	9	0.5
Form	Genre	24	1.3
Form	No subcat.	6	0.3
NRK-jargon	No subcat.	18	1.0
No category		18	1.0
Total		1828	100.0

The majority of tags represent subjects, which is in line with findings from other studies (Golder & Huberman, 2006; Laine-Hernandez & Westman, 2006), but other facets are also judged by the indexers to be important retrieval entry points. Of these names is the most common facet. We have broken down our data to see how tag categories were distributed among the different editorial offices (Table 5).

The most interesting finding in Table 5 is the distribution of tags created by the sports office. We see that it differs from the other offices in tagging far more names (59.5 % of all Sport tags are names) and far less subjects (36.7 %) than the other offices (61.5 %). Apparently, names are important for sport program retrieval whereas subjects are comparably less important. An inspection of sports tags shows many examples of names of individual athletes and teams. In culture and entertainment as well as documentaries and fact programs, the trend seems to be the opposite. Among the names found in the culture and entertainment-tags there are many song titles.

Table 5 Distribution of tag categories among editorial offices

Editorial office	Subject	Place	Name	Time	Form	NRK-jargon	No cat	Total
News	60.9 (461)	12.3 (93)	23.5 (178)	0.9 (7)	0.8 (6)	0.3 (2)	1.3 (10)	100 (757)
Culture and entertainment	67.1 (367)	5.9 (32)	18.3 (100)	1.3 (7)	4 (22)	2.2 (12)	1.3 (7)	100 (547)
Sport	36.7 (104)	1.8 (5)	59.3 (168)	1.8 (5)	-	-	0.4 (1)	100 (283)
Documentaries and fact	77.7 (150)	7.8 (15)	13 (25)	-	1 (2)	0.5 (1)	-	100 (193)
Health, consumer and life style	83.9 (26)	6.5 (2)	-	-	-	9.7 (3)	-	100 (31)
Children	100 (17)	-	-	-	-	-	-	100 (17)
Total	61.5 (1125)	8 (147)	24.8 (471)	1 (19)	1.6 (30)	1 (18)	1 (18)	100 (1828)

RQ2: How does decentralized indexing influence consistency?

A common problem of uncontrolled vocabularies is the lack of consistency in the selection of index terms. We examined how this characterizes our tag collection. In addition, we wanted to see whether the indexing policy of NRK secures consistent exhaustive indexing. We measure indexing exhaustivity by counting the number of tags per program, the hypothesis being that programs with many tags are more exhaustively indexed than those with few (Lancaster, 2003). In addition, our informants supply us with information on their tagging practice.

NRK's indexing policy does not prescribe a specific number of tags per program part, and in the interviews the informants told us that their tagging differs a lot. One informant believed one tag per part was enough, stating, "it gets so messy when you have a lot of tags". Another informant had a

practice of adding three tags per part. These two informants clearly follows their own guidelines rather than the institution's indexing policy.

Our data set consists of 56 programs. The most tagged program, with 84 tags, was a news program in Sami ("Oddasat"), which consisted of 12 parts. Twenty-seven programs had fewer than 30 tags. News programs in general contained the most tags; we have therefore looked at these in more detail to see if we can find any tag patterns.

News programs[5] are on average tagged 43.1 times per program (compared to 32.6 tags for all programs) and 5.3 times per tagged program part (but only 1.88 tags per news program part in total[6]). In Figure 2 we see the tag distribution of tags for all news programs in our sample. The program with most tags, Oddasat (84 tags), is one of the shortest programs in our data set, lasting only 15 minutes. There are six different types of news programs and we cannot see that a consistent pattern exists across the programs. The evening news ("Kveldsnytt"), which is a short program (15 minutes), contained relatively fewer tags on average. The five Sami news programs had 84, 53, 39, 39 and 25 tags respectively. The main daily news program, Dagsrevyen (lasting 45 minutes), had on average 52 tags, which was far more than the other programs. The five editions in our sample had, however, a very inconsistent number of tags, varying between 34 and 75 tags. The morning news (Morgennytt), which is the longest program (2 hours), on the other hand, has a quite stable number of tags, four out of the five shows have between 37 and 44 tags (the fifth have 56 tags assigned to it). Also Dagsrevyen 21 (the 9 o'clock news, lasting 30 minutes) is quite stable, varying between 29 and 46 tags.

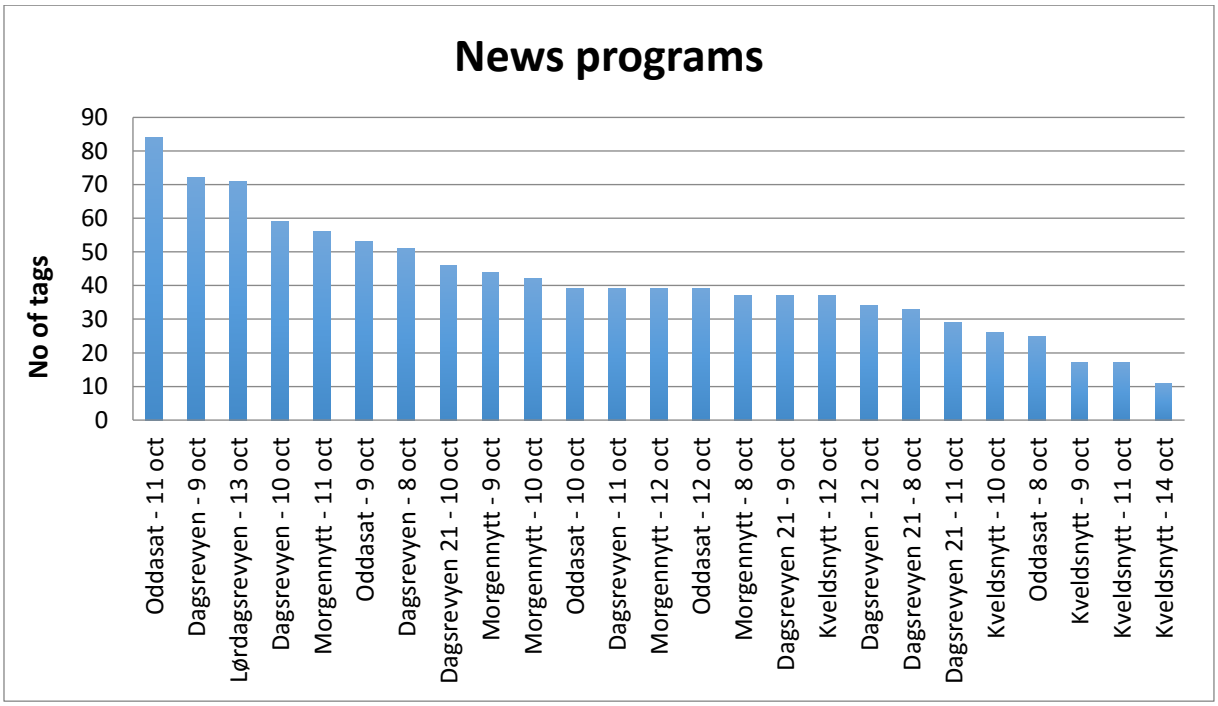

Figure 2 Distribution of tags in News programs

Culture and entertainment programs were on average assigned fewer tags (28 per program), but also this program category had an inconsistent distribution of tags ranging from one to 64 tags. The

[5] Which included 757 tags made by the News office and 279 tags made by Sport, in total 1036 tags
[6] In total there were 553 news program parts, but only 197 of them were tagged

category, however, consists of a larger variety of programs, and programs are not broadcast daily, thus it is difficult to compare them.

The most tagged program part is from Lindmo, a talk show from the Culture and entertainment office. This part contains 23 tags. The tag distribution is between 1 and 23 tags, with an average of 4.75 tags per tagged program part.

It is hard to find a consistent pattern in tagging across program types from the same editorial office. Where we have several editions of the same program (i.e. the news programs) we find consistent number of tags for some programs, but not all. In order to draw more solid conclusion we would need data from a longer period of time.

Analysis of tag frequency (Table 6) shows that more than half the tags (55.5 %) were used only once during the observed week. This suggests that the indexers do not confer with the list of tags previously used and that they prefer to create new tags. One of our informants confirms this, and says that she usually overrules the tag list. An internal evaluation performed by the Metadata section in NRK also shows similar results: they report that 75 % of the tags used for a major news case were used only once.

Table 6 Tag frequency

Frequency	No of tags	Total freq	Percentage
1	1015	1015	55.5 %
2	206	412	22.5 %
3	68	204	11.2 %
4	9	36	2.0 %
5	5	25	1.4 %
6	7	42	2.3 %
7	2	14	0.8 %
8	1	8	0.4 %
9	3	27	1.5 %
10	2	20	1.1 %
11	1	11	0.6 %
14	1	14	0.8 %
Total		1828	100.0 %

We analyzed the tags to identify terminological challenges such as spelling errors, foreign language, plural forms and capital letters.

In all, our sample contains 46 examples of spelling errors, 21 tags in foreign languages (other than Norwegian bokmål, which is the only language allowed according to the guidelines), 123 tags in plural form (the guidelines states all terms should be in singular form) and 333 terms containing capital letters (according to the guidelines all terms, including personal names, should be written in lowercase). This means that 523 out of 1828 tags, 28.5 %, were formulated in conflict with the internal guidelines. In addition, 316 tags consisted of more than one word, many of which are names, e.g. the song title “Det lypte let i den spreke guten”. The guidelines states that preferably tags should consist of only one term. In our material Name and Place tags represent a large majority of the multi-word tags. Thus we consider the guidelines, in general, to be fulfilled in this respect. Our findings are similar to the results of previous research on folksonomy tags (e.g. Guy & Tonkin, 2006). It should

also be noted that some of these problems are easy to solve if the retrieval system is implemented to support word stemming and made case independent.

Our results indicate that NRK's implementation of decentralized indexing results in low indexing consistency. This may be due to the indexers inability or unwillingness to follow the organization's indexing policy. It may be that the training they receive is insufficient, that the software is difficult to use or that the NRK's goals for the goals are unclear. Some of these issues are discussed in the next section.

RQ3: How do indexers' practice and motivation affect indexing quality?

To learn more about how the practical indexing work takes place and affects the quality of tags we asked the informants about issues such as their tagging procedures, how much effort they invested and their motivation.

Our five informants (Table 7) have between half a year and 15 years work experience in NRK and belong to three different editorial offices: News, Entertainment and Children. Two of them register metadata irregularly whereas two do it daily and one at least once per week. Four of the five are directly involved with the TV production process and one has superior responsibility for metadata registration and technical organization in one of the editorial offices. None of the informants are educated in Library and Information Science or had similar formalized metadata background prior to starting in NRK, but three of them tag privately in social web sites. They are clearly representative of NRK's indexers with respect to their high domain knowledge and low professional indexing knowledge.

Table 7 Informants

Informant	Year of birth	Seniority in NRK	Experience with indexing	Indexes
Informant 1	1973	10 years	1 year	Daily
Informant 2	1962	15 years	4 years	Weekly
Informant 3	1984	5 months	5 months	Daily
Informant 4	1979	11 years	2.5 years	Rarer than monthly
Informant 5	1979	5 years	1.5 years	Rarer than monthly

The informants tag programs that they in a varying degree are involved in the production of. In our group, there were no journalists or reporters, but two of the informants, whose additional responsibility is to check their colleagues' indexing, tell that they feel a lot of time is spent "nagging" on journalists to add tags. They say that many of the employees feel tagging is considered a duty that is forced upon them. For this reason, NRK is trying to make their employees understand the potential of tagging for internal as well as external retrieval, e.g. by showing the effect the tagged TV programs can have on ratings.

The time spent on tagging varies depending on the length of the program and its individual parts. Two informants say they spend approximately 45 minutes per program adding metadata, including tags. The two other informants regularly performing indexing use "from five minutes per item and upwards" and "approximately 20 minutes per program" respectively. The fifth informant have a responsibility for overseeing and teach his colleagues how to index, but do not index regularly.

The indexing rules states that the tags should be derived from the programs and individual program parts. This is in accordance with Hulme's (1911) concept "literary warrant". We asked the informants what their focus was at the moment of indexing: if they analyzed images, sound and/or the whole

event taking place. The answers we got differed; one said he used to read the manuscript or other texts about the program and look through the item, another stated “I think about the item as a whole, I think about the images, sounds and persons participating. And I think ‘situations’”. A third informant told us that she was not consistent in her approach to indexing: sometimes she focused on the images, other times on the activities taking place.

According to the guidelines, the indexer should add “common synonyms”. The purpose being to increase the number of access points for program retrieval. Only one of our informants, however, stated that he consciously added synonyms during indexing.

The list of tags, our semi-controlled vocabulary, is used to a varying degree by the informants. Tags that match the indexer’s input are automatically suggested as index terms. One claims to use it actively in order to secure consistency in her tagging whereas another informant finds it “irritating” and that she “chooses to overlook the suggestions” in favor of creating her own terms.

When asked about the target groups of their tagging, two informants stated that they primarily had the external audience in mind. One informant said that she was very aware of using “an easily understandable language” in order to secure that children and non-native speakers could get access to the programs. Another informant agreed and said that other metadata fields were used by the internal staff for retrieval. A third informant told us that his focus is on describing the program for someone “who doesn’t know the content”, independent of them being internal or external. The informant with superior responsibility for metadata registration said that it was difficult for the taggers to avoid internal terminology. In our data, however, we found very few tags representing jargon (approximately 1 % of all tags).

Conclusion and discussion

We have performed an analysis of tagging of television programs performed by the production staff in a public broadcaster. We have called this decentralized indexing. The aim was to examine how decentralized indexing with the use of a semi-controlled tag vocabulary influenced the indexing quality. We have investigated three specific research questions: RQ1) What characteristics of television programs are indexed? RQ2) How do decentralized indexing influence indexing consistency? RQ3) How do indexers’ practice and motivation affect indexing quality?

Concerning RQ1 we found that the majority of tags used to a large degree (61.5 %) represent subjects whereas 25.8 % represent names. This differs across editorial offices. In particular, it is interesting to note that the Sport office has an overrepresentation of Name tags compared to the other programs. We found no examples of “personal” (Kipp, 2007) or “opinion expression” (Marlow et al., 2006) tags in our material, which is a sign that the indexing guidelines are followed and that the vocabulary is under some control. Our findings are also consistent with Golder and Huberman (2006), who found that the majority of tags are topical.

With respect to tagging consistency (RQ2) we have observed several critical issues: the distribution of tags per program differs a lot, which results in varying indexing exhaustivity; a lot of the terms are used only once; and several terminological errors appear, such as spelling errors and incorrect use of grammatical form. Thus the metadata are quite “messy” (Smith, 2008). Most important, however, is the finding that very few programs are tagged at all. Only 56 out of 200 (249 including reruns), i.e. only 28 % of all programs were tagged and of the 823 program parts in these 56 programs only 371 (45 %) included tags. This is a clear indication that decentralized indexing can also cause low indexing coverage. Low coverage does not equal low consistency, but both result in poorer program retrieval.

For RQ3, among the five indexers that served as our informants, we found different practices with respect to their use of the list of used tags (i.e. the 'semi-structured vocabulary'). Indexing differed from conscious use to conscious non-use of tags. The indexers used different aspects of a program to determine its content and their perception of target groups for the tags also differed. We also got the impression that there was low motivation for tagging among the production staff. Our informants spent short time indexing a program, in particular when compared with professional indexers in BBC (Dowman et al., 2005).

Of course, our five informants are not necessarily representative of all personnel involved in tagging at NRK. On the other hand, they represent a varied group of people with respect to their background and practices and have in common that they are domain experts in television production and are not professionally trained indexers. This is also true for the production staff in NRK in general. We therefore believe our findings tell a true story about tagging practices in a company with decentralized indexing. It would have been interesting to compare the current practice with previous indexing in NRK, when it was centralized and done by professional indexers.

Based on our findings we draw the conclusion that the use of indexers with high domain knowledge in television production but low indexing knowledge negatively affects indexing quality. In addition, the use of a semi-controlled vocabulary probably further reduces indexing quality. We find that indexing with a vocabulary consisting of user created terms under some control by the organization's archive and research department is reminiscent to indexing with a non-controlled vocabulary ('folksonomy'), although we do not find any use of "personal" tags. We have not analyzed the specificity of the terms used, but have found that indexing exhaustivity differs a lot. We find low tagging consistency in our material. Low consistency is to be expected with an uncontrolled vocabulary and our findings suggest this is also the case when the vocabulary is semi-controlled. A vocabulary with high term specificity also is known to cause low consistency, but without a systematic analysis of NRK's vocabulary, we cannot conclude that this is the case here.

A semi-controlled vocabulary lies somewhere between a controlled vocabulary and a folksonomy. The indexers are free to choose terms to describe the resource's content, but must adhere to a set of guidelines stating what aspects to describe, the form terms should take and the explicit use of synonyms.

We have investigated a case where the institution practices decentralized indexing using a semi-controlled vocabulary. Together with term suggestions from the list of tags, the guidelines do, however, not result in good indexing practice. We believe that domain knowledge does not compensate for the lack of indexing training and motivation. Thus the costs saved in not using professional indexers may be transferred to the end-users who will compensate for bad indexing by spending more time on queries.

Future research could investigate the match of queries and semi-controlled index terms through analysis of query transaction logs. It would also be interesting to compare how different groups of indexers use semi-controlled vocabularies. A third way to follow up our work would be to assess the use of automatic term assignment, based on manuscripts or subtitles (almost all NRK programs are subtitled), as an alternative way to index the programs.

References

Agius, H., Angelides, M. C., & Zad, D. D. (2012). Experimenting with tagging and context for collaborative MPEG-7 metadata. *Multimedia Tools and Applications*, *62*(1), 143–177. http://doi.org/10.1007/s11042-011-0984-x

Ames, M., & Naaman, M. (2007). Why We Tag: Motivations for Annotation in Mobile and Online Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 971–980). New York, NY, USA: ACM. http://doi.org/10.1145/1240624.1240772

ARD/ORF/ZDF. (2008). *Regelwerk Mediendokumentation: Fernsehen - Richtlinien für die Formalbeschreibung und Inhaltserschließung von Fernsehproduktionen* (Version 1.0). Retrieved from http://www.bui.haw-hamburg.de/pers/ulrike.spree/praesentation/regelwerk_fernsehen_komplett.pdf

Bogers, T., & Petras, V. (2015). Tagging vs. Controlled Vocabulary: Which is More Helpful for Book Search? Retrieved from https://www.ideals.illinois.edu/handle/2142/73673

Cleverdon, C. W., & Keen, M. (1966). *Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 2, Test results* (Technical Report).

Dowman, M., Tablan, V., Cunningham, H., & Popov, B. (2005). Web-assisted Annotation, Semantic Indexing and Search of Television and Radio News. In *Proceedings of the 14th International Conference on World Wide Web* (pp. 225–234). New York, NY, USA: ACM. http://doi.org/10.1145/1060745.1060781

Ekenel, H. K., & Semela, T. (2011). Multimodal genre classification of TV programs and YouTube videos. *Multimedia Tools and Applications*, *63*(2), 547–567. http://doi.org/10.1007/s11042-011-0923-x

Geisler, G., Willard, G., & Ovalle, C. (2011). A Crowdsourcing Framework for the Production and Use of Film and Television Data. *New Rev. Hypermedia Multimedia*, *17*(1), 73–97. http://doi.org/10.1080/13614568.2011.552645

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198–208. http://doi.org/10.1177/0165551506062337

Goodrum, A. (2000). Image Information Retrieval: An Overview of Current Research. *Informing Science*, *3*, 2000.

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up Tags? *D-Lib Magazine*, *12*(1). http://doi.org/10.1045/january2006-guy

Hulme, E. W. (1911). Principles of Book Classification. *Library Association Record*, (Dec.), 444–449.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. http://doi.org/10.1108/eb026526

Kipp, M. E. I. (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.

Kipp, M. E. I. (2007). @toread and Cool: Subjective, affective, and associative factors in tagging. *Canadian Journal of Information and Library Science-Revue Canadienne Des Sciences De L Information Et De Bibliotheconomie*, *31*(3-4), 274–274.

Kipp, M. E. I. (2011a). Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. *Knowledge Organization*, *38*(3), 245–261.

Kipp, M. E. I. (2011b). User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike / Le contexte de l'indexation des usagers, des créateurs et des professionnels : une exploration des pratiques d'étiquetage social sur CiteULike. *Canadian Journal of Information and Library Science*, *35*(1), 17–48. http://doi.org/10.1353/ils.2011.0008

Kipp, M. E. I., & Campbell, D. G. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–18. http://doi.org/10.1002/meet.14504301178

Knautz, K., & Stock, W. G. (2011). Collective indexing of emotions in videos. *Journal of Documentation*, *67*(6), 975–994. http://doi.org/10.1108/00220411111183555

Kvale, S. (2007). *Doing Interviews*. SAGE Publications.

Laine-Hernandez, M., & Westman, S. (2006). Image semantics in the description and categorization of journalistic photographs. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–25. http://doi.org/10.1002/meet.1450430148

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

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *Acm Transactions on Multimedia Computing Communications and Applications*, *2*(1), 1–19. http://doi.org/10.1145/1126004.1126005

Markkula, M., & Sormunen, E. (2000). End-User Searching Challenges Indexing Practices Inthe Digital Newspaper Photo Archive. *Information Retrieval*, *1*(4), 259–285. http://doi.org/10.1023/A:1009995816485

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (pp. 31–40). New York, NY, USA: ACM. http://doi.org/10.1145/1149941.1149949

Montagnuolo, M., & Messina, A. (2007). Automatic Genre Classification of TV Programmes Using Gaussian Mixture Models and Neural Networks. In *18th International Workshop on Database and Expert Systems Applications, 2007. DEXA '07* (pp. 99–103). http://doi.org/10.1109/DEXA.2007.92

Nov, O., Naaman, M., & Ye, C. (2008). What Drives Content Tagging: The Case of Photos on Flickr. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1097–1100). New York, NY, USA: ACM. http://doi.org/10.1145/1357054.1357225

Peters, I. (2009). *Folksonomies: indexing and retrieval in Web 2.0*. Berlin: De Gruyter ; Saur.

Rocha, R., & Cobo, Á. (2011). Feature selection strategies for automated classification of digital media content. *Journal of Information Science*, *37*(4), 418–428. http://doi.org/10.1177/0165551511412028

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, *4*(4), 247–375. http://doi.org/10.1561/1500000009

Sa, N., & Yuan, X. (2013). What Motivates People Use Social Tagging. In A. A. Ozok & P. Zaphiris (Eds.), *Online Communities and Social Computing* (pp. 86–93). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-39371-6_10

Smeaton, A. F. (2004). Indexing, Browsing, and Searching of Digital Video. *Annual Review of Information Science and Technology (ARIST)*, *38*, 371–407.

Smith, G. (2008). *Tagging: People-powered Metadata for the Social Web* (1st ed.). Thousand Oaks, CA, USA: New Riders Publishing.

Spiteri, L. F. (2007). The Structure and Form of Folksonomy Tags: The Road to the Public Library Catalog. *Information Technology and Libraries*, *26*(3), 13–25. http://doi.org/10.6017/ital.v26i3.3272

Thomas, M., Caudle, D. M., & Schmitz, C. M. (2009). To tag or not to tag? *Library Hi Tech*, *27*(3), 411–434. http://doi.org/10.1108/07378830910988540

Turner, J. M. (1995). Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: research results. In *Proceedings of the 58th Annual Meeting of the American Society for Information Science*. Chicago, Ill.: ASIS. Retrieved from http://www.asis.org/asis-95/papers/turner.html

Wal, T. V. (2007, February 2). Folksonomy Coinage and Definition. Retrieved from http://vanderwal.net/folksonomy.html