# Ämnesord, redundans, nationalbibliografi

## ämnesordens redundans i nationalbibliografin

**Edvin Lindström**

Examensarbete (30 högskolepoäng) i biblioteks- och informationsvetenskap för masterexamen inom ABM-masterprogrammet vid Lunds universitet.
Handledare: Jonas Fransson
År: 2017

## Title

Subject headings, redundancy, national bibliography: redundancy of subject headings in the Swedish national bibliography

## Abstract

There is an ongoing discussion in information retrieval research regarding the merits of subject indexing with controlled vocabularies as opposed to relying on free-text descriptions. Earlier research has compared subject headings with other metadata, such as document titles or tables of contents, to determine the relative value of subject headings as subject access points in an information searching process. If a subject heading associated with a document is also found for example in the document's title, it can be seen as redundant – adding little value to the description of the document, and not making the document easier to find.

This study compares subject headings in a subset of the Swedish national bibliography with document titles and tables of contents, in order to find whether certain kinds of subject headings more often than others tend to be redundant. Earlier research has focused on subject headings as a whole, without considering that different kinds of subject headings may have different properties. Several ways to distinguish ”kinds” of subject headings are proposed.

The main findings are that personal names as subject headings tend almost always to be redundant; subject headings specifying a time period are very rarely redundant; and geographical terms show a strong tendency for specific place names to be redundant, whereas names of countries and provinces are less so. There is also a tendency for longer terms to be less redundant than short ones.

These findings can all be explained individually, but there is no one single conclusion to draw from the results. It's important to bear in mind that a subject heading being strictly speaking redundant does not mean that it has no value, since subject headings also serve purposes such as connecting documents on the same topic. Findings like these make for a more nuanced debate for and against the use of controlled vocabularies, and may be of use when designing support for subject headings in library catalogs.

Master's thesis

**Keywords**

ALM, library and information studies, subject headings, subject indexing, controlled vocabularies

**Nyckelord**

ABM, biblioteks- och informationsvetenskap, ämnesord, ämnesindexering, kontrollerade vokabulärer

# Innehåll

# Tabeller



# Figurer

*”Att slå upp”, säger gubbfan, ”är oftast det bästa.”*
*Han är inte så ung att han vet det mesta.*

Alf Henrikson

# 1 Inledning

Bibliotekskatalogers användare vill kunna hitta vad de letar efter. Ofta letar de inte efter ett specifikt dokument, utan dokument som behandlar ett visst ämne. För att göra det möjligt att hitta alla dokument som behandlar samma eller liknande ting, bör bibliotekets samlingar på något sätt ordnas efter ämne. Ett sätt att göra detta är genom klassificering, det vill säga att ge varje dokument en viss kod som motsvarar dess placering i vetandets universum. Ett annat sätt är att ämnesindexera: att med en eller flera allmänspråkliga termer beskriva dokumentets huvudsakliga innehåll.

Traditionell indexering använder sig av en kontrollerad vokabulär, alltså en påbjuden samling termer som vanligtvis är ordnade efter deras inbördes hierarkiska relationer. Termerna kallas *ämnesord*. Fördelarna gentemot fri indexering med vilka ord som helst är bland annat att samma sak alltid benämns med samma term, och att den hierarkiska strukturen i vissa kataloger gör det möjligt att navigera mellan besläktade ämnen.

Men dokumenten behöver kanske inga ämnesord. Det finns ju ofta annan information om dem som också ger ledtrådar om ämnet: titeln brukar väl vara informativ, och innehållsförteckningar och korta sammanfattningar bidrar ytterligare. Dessutom utvecklar man nya metoder för exempelvis automatisk indexering och användartaggning. Det finns en strömning inom biblioteks- och informationsvetenskapen som vill ersätta ämnesorden med nya tekniker eller slopa dem helt.

Emot denna strömning finns de som menar att ämnesorden fortfarande behövs, och argumenterar för detta bland annat genom att visa hur många fler relevanta dokument man hittar när katalogposterna har ämnesord än när de saknas. Forskningen tittar nästan alltid på ämnesorden som helhet: si och så många procent av de relevanta dokumenten hade inte hittats utan ämnesord, så här mycket bättre blir ämnesbeskrivningarna tack vare ämnesorden.

Det saknas forskning som i närmare detalj studerar hur fördelningen av ämnesord ser ut i en katalog, och som undersöker om vissa ämnesord gör mer nytta än andra. Diskussionen om ämnesordens framtid kan föras på en djupare nivå om man känner till vilka typer av ämnesord som tenderar att göra dokument lättare att hitta, och vilka som i praktiken inte tillför särskilt mycket. Det är här som denna uppsats kommer in.

## 1.1 Syfte och frågeställningar

Syftet med den här uppsatsen är att skapa ökad kunskap om vilka egenskaper de ämnesord har som tenderar att göra dokument mer sökbara. Ett annat sätt att uttrycka detta är att uppsatsen undersöker vilka typer av ämnesord som tillför dokumenten unika *subject access points*. Liksom Hjørland och Kyllesbech Nielsen (2001) förstår jag med ”subject access point” (SAP) varje element i en bibliografisk post som tillåter användaren att nå

fram till det beskrivna dokumentet genom en sökning på ett visst ämne. Ett ämnesord är en typisk SAP, ja själva dess främsta syfte är att fungera som sådan: genom en sökning på ”kaffe” hittas *Fikaboken : små råd och recept för stora fikaälskare* tack vare att dess katalogpost tillfogats ämnesordet ”Kaffe”. Sökordet återfinns ingen annanstans i posten, så ämnesordet skapar en unik SAP. Boken *Kaffe, choklad, tango* har samma ämnesord, men en sökning på ”kaffe” hade hittat den ändå, eftersom ordet ingår i titeln. Ämnesordet skapar ingen unik SAP. Med andra ord: vi ser att ämnesordet förbättrade sökbarheten för den första boken, men inte för den andra.

I uppsatsen undersöks förhållandet mellan ämnesord å ena sidan, och titel respektive innehållsförteckning å den andra, i ett urval katalogposter i den svenska nationalbibliografin. Frågorna som ställs är:

- Vilka typer av ämnesord tenderar att återfinnas också i titeln eller innehållsförteckningen?
- Vilka typer av ämnesord tenderar att *inte* återfinnas i titeln eller innehållsförteckningen, och därmed skapa unika SAP:er?

Begreppet ”typer av ämnesord” diskuteras närmare i avsnittet om undersökningens metod. Med utgångspunkt i svaren på frågorna ovan vill undersökningen vidare besvara frågorna:

- Vilka är orsakerna till de funna mönstren i ämnesordens redundans?
- Vad betyder ämnesordens redundansmönster för deras nytta?

Termen ”redundans” innefattar inget värdeomdöme, utan ska förstås som en rent deskriptiv benämning på det förhållandet att ett ämnesord återfinns i en titel eller innehållsförteckning.

Undersökningen begränsar sig till ämnesord ur den kontrollerade vokabulären Svenska ämnesord. Detta är den största och mest använda ämnesordlistan på svenska. En bieffekt av och ett sidosyfte med uppsatsen är att bidra till forskningen om denna understuderade ämnesordslista.

## 1.2 Disposition

Denna inledning följs av ett avsnitt 2 som närmare presenterar ämnesindexeringen ur både praktiskt och teoretiskt hänseende, där också Svenska ämnesords historia och uppbyggnad har sin plats. Avsnitt 3 presenterar tidigare forskning om ämnesindexering, samt det viktigaste som publicerats om Svenska ämnesord. I avsnitt 4 relateras studien till ett teoretiskt ramverk. Här introduceras också ett analytiskt verktyg vars tillämpning vidare diskuteras i avsnitt 5, som redogör för undersökningens metod och metodologiska överväganden. I avsnitt 6 redovisas studiens resultat, och för överskådlighetens skull analyseras och förklaras dessa i anslutning till sin presentation. Ett avslutande avsnitt 7 lyfter så blicken och diskuterar i vilken mån forskningsfrågorna har fått svar.

# 2 Bakgrund

I det här avsnittet beskrivs först vad indexering är, dess syften, hur den går till och några viktiga begrepp. Därefter presenteras Svenska ämnesord, dess tillkomst och uppbyggnad.

## 2.1 Indexering

Ämnesindexeringen[1] är en del av den bibliografiska beskrivningen av ett dokument med hjälp av metadata, närmare bestämt metadata som beskriver dokumentets ämne eller ämnen – med andra ord dess innehåll, eller vad det handlar om.[2] Indexeringen kan ske antingen med fritt valda termer, eller (oftare) med termer ur en kontrollerad vokabulär, vilket också är den typ av indexering som den här uppsatsen kommer att studera. En kontrollerad vokabulär är en lista över de godkända termer som får användas, i regel med någon form av semantisk struktur. Strukturen är med Lancasters (1991:14) av mig fritt försvenskade ord och exempel ägnad att:

1. kontrollera synonymer genom att välja en form som standard och hänvisa till den från alla andra;
2. skilja mellan homografer. Till exempel är ”Dockor (fartygsbassänger)” något helt annat än ”Dockor (leksaker)”; och
3. sammanföra eller länka mellan de termer vars betydelser är närmast besläktade. Två slags förhållanden kan uttryckligen identifieras: det hierarkiska och det icke-hierarkiska (eller *associativa*) förhållandet. Termen ”Gifta kvinnor” är hierarkiskt besläktad med ”Kvinnor” åt ena hållet och ”Hemmafruar” åt det andra, och därtill associativt besläktad med termer som ”Äktenskap” och ”Familjer”, vilka inryms i andra hierarkier.

Kontrollerade vokabulärer kan delas in i några olika kategorier. Taylor och Joudrey (2009:334ff) skiljer mellan *ämnesordlistor*, *tesaurer* och *ontologier*. Tesaurer skiljer sig från ämnesordlistor i att deras termer alla är atomära av typen ”Fiskar”, medan ämnesordlistor kan innehålla kombinerade termer som ”Fiskar i litteraturen”. Tesaurer är som en konsekvens mer strikt hierarkiska, med oftast bara en överordnad term per term. Ontologier liknar båda dessa, men väljer inte ut en term som standard bland synonymer. Lancaster (1991) utelämnar ontologierna men räknar klassifikationssystem som en typ av kontrollerad vokabulär. Dessa är ganska annorlunda i jämförelse, som till exempel använder koder som ”Ugi” eller ”597” istället för språkets vanliga ord. De kommer fortsättningsvis i den här uppsatsen inte att inräknas i begreppet kontrollerad vokabulär. Svenska ämnesord, ämnet för uppsatsens undersökning, är som tidigare nämnts en ämnesordlista (vartill namnet ju är en god ledtråd).

[1] För enkelhets skull används termerna ”ämnesindexering” och ”indexering” i den här uppsatsen omväxlande utan betydelseskillnad.

[2] ”Ämne”, ”innehåll” och vad något ”handlar om” är inte nödvändigtvis helt synonyma uttryck i alla sammanhang, men det ligger utanför den här uppsatsens intressen att utreda skillnaderna. För diskussion om detta, se exempelvis Hjørland (1993) och Lancaster (1991:10–13).

Indexering kan vara i olika utsträckning prekoordinerande eller postkoordinerande. Prekoordinering innebär att relevanta begrepp sammanförs redan vid indexeringstillfället. En bibliografi över rökningens amerikanska 1900-talshistoria kan få den prekoordinerade ämnesordssträngen ”Rökning – historia – Förenta staterna – 1900-talet – bibliografi”. Med postkoordinerande indexering delas termerna istället upp i ”Rökning”, ”Historia”, ”Förenta staterna”, ”1900-talet” och ”Bibliografi”. Exempel på mer subtil prekoordinering är samordning av begrepp så som ”Kvinnor och politik” (istället för ”Kvinnor” och ”Politik”) eller ”Svensk litteratur” (inte ”Litteratur” och ”Sverige”).[3] Gartner (2016:69) beskriver prekoordinering som ett uttrycksfullt sätt att skapa meningsfulla semantiska enheter, men som förutsätter att indexeraren kan göra en riktig bedömning av de atomära begreppens inbördes sammanhang.

Indexeringens syften hänger samman med den bibliografiska beskrivningens syften överhuvudtaget. Den första som uttryckligen formulerade de sistnämnda var Charles A. Cutter år 1876. Han menade att en bibliotekskatalog ska:

1. hjälpa någon hitta en bok vars författare, titel eller ämne är känt,
2. visa vad biblioteket har av en viss författare, inom ett visst ämne eller inom ett visst slags litteratur, samt
3. vara till stöd vid valet av bok med avseende på upplaga eller karaktär.

(Cutter 1876a se Svenonius 2000:15; min översättning)

I modern tid har följande krav på bibliografiska beskrivningar formulerats på uppdrag av IFLA (1998/2006:65):

- att hitta material som motsvarar de sökkriterier användaren angivit (det vill säga att genom en sökning som använder ett attribut eller en relation som hänger samman med entiteten lokalisera en eller flera entiteter i en katalog eller databas);
- att identifiera en entitet (till exempel att försäkra sig om att ett beskrivet dokument verkligen är det sökta, eller att skilja mellan två eller flera entiteter med liknande karakteristika);
- att välja en entitet som är anpassad till användarens behov (till exempel att välja en entitet som motsvarar användarens behov vad beträffar innehåll, fysiskt format etc., eller att välja bort en entitet som inte gör det);
- att anskaffa eller få tillgång till den entitet som beskrivs (det vill säga att få tag i en entitet genom inköp, lån, etc. eller få tillgång till en entitet elektroniskt via nätuppkoppling till en fjärrdator).[4]

[3] Exemplen är hämtade från Kungliga biblioteket (u.å.:37–38).

[4] En uppenbar skillnad mellan formuleringarna är att Cutter talar om ”böcker” medan IFLA talar om ”entiteter”. I den här uppsatsen använder jag av bekvämlighet termen ”dokument” för det som beskrivs av en katalogpost, vilket inte ska tolkas som ett ontologiskt ställningstagande om vad som kan och inte kan utgöra ett dokument.

Målen har som synes inte förändrats mycket på ett och ett halvt sekel. IFLA:s lista är i stort en moderniserad version av Cutters, med tillägg av den sista punkten om tillgång. För alla punkter utom denna spelar ämnesindexeringen en roll. Ett ämnesindexerat dokument är lättare att *hitta*, eftersom det kan sökas fram på ämne. Det är lättare att *identifiera*, eftersom ämnesorden kan bekräfta att det handlar om det som eftersöks. Indexeringen gör det också lättare att *välja* dokument utifrån deras innehåll. Ämnesindexeringens viktigaste funktion är nog emellertid den första – att förbättra möjligheterna att hitta dokument – och det är ur detta perspektiv indexering nästan alltid diskuteras. Den förbättrar mer specifikt ämnesåtkomsten (eng. *subject access*), det vill säga förenklar sökning efter dokument inom ett visst ämne. På kortkatalogens ännu inte bortglömda tid var ämnesorden den enda ingången till ämnessökningar, men i dagens digitala kataloger som tillåter fria sökningar i alla metadatafält samtidigt kan i princip varje litet ord i en katalogpost skapa en sökträff på ämne (Olson och Boll 2001:2). Det är detta som har gjort vikten av ämnesindexering möjlig att ifrågasätta på de sätt som ska diskuteras mer ingående i forskningsöversikten, och det är inom denna kontext som den här uppsatsen kommit till.

ISO-standard 5963:1985 ger en beskrivning av indexeringsprocessen i tre steg (ISO 1985):

1. Undersökning av dokumentet för att bestämma dess ämnesinnehåll.
2. Konceptuell analys för att avgöra vilka aspekter av innehållet som bör representeras.
3. Översättning av dessa aspekter eller begrepp till en kontrollerad vokabulär.

Det går inte alltid att skilja de tre stegen tydligt åt. Lancaster (1991:8) väver ihop de första två stegen i ett som han kallar ”konceptuell analys”, och menar att denna i praktiken kan ske samtidigt med översättningen även om det rör sig om två förnuftsmässigt åtskilda processer. Hjørland (1993:40f) menar på liknande sätt att ämnesanalysen ofta påverkas av den vokabulär som ämnena ska uttryckas i, och ser både positiva och negativa effekter av detta: å ena sidan kan vokabulärens begränsningar och inbyggda teoretiska perspektiv tvinga fram en viss tolkning av dokumentet, men å andra sidan effektiviseras ämnesanalysen av att inte göras mer utförlig än den tänkta vokabulären ändå kan uttrycka.

ISO-standardens första steg, identifieringen av ämnesinnehållet, kan gå till på olika sätt. En naturlig utgångspunkt är titeln, vilket har den uppenbara bristen att ett dokuments titel kan vara missvisande, retoriskt tillspetsad eller poetiskt omskrivande. Hjørland (1993:66) vill till och med mena att ett dokument inte nödvändigtvis handlar om det som dess författare tror, varför den mest oskyldigt uppriktiga titel kan vara ett villospår. Detsamma gäller förstås också andra uttryck för författarens uppfattning om sitt verk. En bättre analys kräver att indexeraren anstränger sig för att utröna dokumentets implicita ämnesinnehåll. Hänsyn bör också tas till de avsedda kataloganvändarnas behov och intressen (Olson och Boll 2001:89; Lancaster 1991:8). Detta kan göras först i standardens andra steg. En vanlig skillnad att göra är den mellan ”document-oriented” och ”request-oriented” indexering. Indexeraren frågar sig i det senare fallet inte bara vad

dokumentet handlar om, utan också med vilka termer användare kan förvänta sig hitta det (Soergel 1985:230). Här snuddas ju också det sista steget, översättning, vid.

Avgörande egenskaper för indexeringens kvalitet är dess *riktighet*, *uttömmandegrad*, *specificitet* och *konsekvens* (Olson och Boll 2001:88–105).[5] Ett sätt att mäta riktigheten är att jämföra indexeringen med de regler och riktlinjer som gäller för den kontrollerade vokabulär som används. Soergel (1994:593f) använder uttrycket ”correctness” och definierar det som frånvaron av felaktiga ämnesord och närvaron av alla lämpliga. Han påpekar att detta måste bedömas utifrån systemets regler, exempelvis den önskade uttömmandegrad som eventuellt stadgas.

Uttömmandegraden mäter hur stor andel av alla relevanta ämnen som indexeras. Olson och Boll (2001) liknar det vid indexeringens bredd. Den hör nära samman med specificiteten, som på motsvarande vis kan ses som indexeringens djup, eller den hierarkiska nivå på vilken de representerade ämnena uttrycks. För att spinna vidare på ett exempel från Olson och Boll: utömmandegraden avgör om indexeringen av en bok som bara kortfattat nämner toypudlar alls ska representera det ämnet, medan specificiteten bestämmer om termen som väljs är ”Toypudlar”, ”Pudlar” eller ”Hundar”. Uttömmandegraden är vanligen en produkt av indexeringens riktlinjer. Specificiteten är dels inneboende i vokabulären – det går inte att indexera med ”Toypudlar” om termen inte finns – och dels ett val som görs vid indexeringen. Enligt en traditionell princip ska ett ämne alltid indexeras med den mest specifika term som är möjlig (Lancaster 1991:26).

Konsekvens innebär att samma ämne genomgående indexeras med samma term, både av samma och mellan olika indexerare. Som Soergel (1994:594) påpekar är detta inget självändamål, eftersom indexeringen kan vara konsekvent felaktig. Konsekvens är däremot en förutsättning för riktighet. Två faktorer inverkar på indexeringens konsekvens enligt Olson och Boll (2001): ju högre uttömmandegrad, desto större sannolikhet att olika indexerare uppfattar de perifera ämnena olika, och ju större vokabulär som används, desto större sannolikhet att de väljer olika termer.

Dessa egenskaper hos indexeringen påverkar sökningarnas *recall* (hur stor andel av alla relevanta dokument som återvinns) och *precision* (hur stor andel av alla återvunna dokument som är relevanta). Dessa mått brukar ofta betraktas som omvänt proportionerliga, det vill säga att det ena minskar när det andra ökar, även om förhållandet inte är fullt så simpelt (Rowley och Hartley 2008:295). Tillfredsställande recall är beroende av indexeringens kvalitet, vilken i sin tur påverkas av dess specificitet och konsekvens. Högre uttömmandegrad ökar vanligen recall på bekostnad av precision. Som Soergel (1994) visar bör begreppen recall och precision användas med varsamhet, eftersom de är avhängiga användarens svårmätbara relevansbedömning av dokumenten.

## 2.2 Svenska ämnesord

Svenska ämnesords tillkomsthistoria beskrivs av Miriam Nauri och Magdalena Svanberg i häftet *Svenska ämnesord: en introduktion* (2004), som refereras i nästa stycke.

Ämnesindexering har länge förekommit på svenska bibliotek, men ursprungligen fanns

[5] *Accuracy*, *exhaustivity*, *specificity* respektive *consistency* på engelska.

ingen gemensam kontrollerad vokabulär. Katalogisatören kunde välja ämnesord tämligen fritt efter eget huvud, utom vid de stora biblioteken som hade egna listor, till exempel KBslagord på KB. Dessa listor saknade emellertid auktoritetsposter och riktlinjer för användning. Arbetet med framtagandet av det som skulle bli Svenska ämnesord tog fart vid 1990-talets början. För att förenkla arbetet utgick man från termerna i det alfabetiska indexet till det väletablerade klassifikationssystemet för svenska bibliotek (SAB). För att kunna dra fördel av internationellt samarbete lutade man sig mot de amerikanska Library of Congress Subject Headings (LCSH): riktlinjerna för Svenska ämnesord modellerades efter riktlinjerna till LCSH, och de svenska ämnesorden översattes till sina motsvarigheter i LCSH. Denna mappning finns kvar och upprätthålls fortfarande kontinuerligt (Kungliga biblioteket 2017a). Databasen Svenska ämnesord skapades i december 1999, och 2002[6] var riktlinjerna klara. I arbetet som ledde fram till det nationella systemet deltog jämte Kungliga biblioteket (KB) bland andra en kommitté under Sveriges allmänna biblioteksförening samt biblioteken vid Stockholms, Uppsala och Göteborgs universitet. Idag underhålls systemet av en redaktion vid KB (Kungliga biblioteket 2017b). En separat lista för indexering av skönlitteratur låg tidigare under Svensk biblioteksförenings ansvar, men överfördes 2014 till redaktionen för Svenska ämnesord; dessa ämnesord är sedan 2016 sökbara i Svenska ämnesords databas, och deras riktlinjer väntas under 2017 föras in i samma riktlinjedokument som de facklitterära ämnesorden (Kungliga biblioteket 2017c).

Svenska ämnesord innehåller omkring 38 000 termer och är särskilt lämpat för ämnen inom humaniora och samhällsvetenskap (Kungliga biblioteket 2017b). I riktlinjerna för Svenska ämnesord (Kungliga biblioteket 2017a:75–90) beskrivs principerna för ämnesordssystemets uppbyggnad. Termers inkludering i listan är dokumentbaserad, det vill säga att det måste finnas minst ett verk som behandlar ämnet i någon databas som använder Svenska ämnesord. Ämnesorden ska vara specifika, men på en nivå som motsvarar användarnas behov och som inte skapar risk för sammanblandning av näraliggande ämnen, varför exempelvis ”Lador” föredras framför ”Hölador” och ”Ängslador”. Föråldrade termer som ”Ordblindhet” undviks till förmån för aktuella som ”Dyslexi”, och svenska termer föredras framför lånord om inte det senare är mer etablerat, som ”Cancer” istället för ”Kräfta”. Eftersom Svenska ämnesord har en bred målgrupp undviks i allmänhet facktermer, vilket innebär att det i listan heter ”Sömngång” och inte ”Somnambulism”. Stavningen följer i första hand Svenska akademiens ordlista, och förkortningar skrivs helst ut, om inte den förkortade termen är välkänd och i allmänt bruk, till exempel ”Aids”. Ämnesorden är substantiv eller substantivfraser och står vanligen i obestämd form, men abstrakta begrepp och specifika företeelser kan stå i bestämd form (”Barndomen”, ”Månen”). Ord som kan skrivas i plural får pluralform (”Aktier”), medan övriga skrivs i singular (”Akrobatik”); ibland används båda i skilda betydelser (råvaran ”Fisk” kontra levande ”Fiskar”). För att skilja mellan ord med samma form men olika betydelse används särskiljande parentestillägg, till exempel ”Stamböcker (minnesböcker)” och ”Stamböcker (register över avelsdjur)”. Synonymkontrollen upprätthålls genom så kallade se-hänvisningar från termer som inte används till deras tillåtna motsvarigheter eller nästan-motsvarigheter, så som ”Ikonologi” hänvisar till ”Ikonografi”, ”Fetma” till ”Övervikt” och ”DJ:s” till ”Diskjockyer”. De närmast över- och under-

[6] Viktoria Lundborg, bibliotekarie Kungliga biblioteket och redaktör Svenska ämnesord, personlig kommunikation 2017-02-13.

ordnade termerna i den hierarki eller de hierarkier som en term ingår i redovisas. För att länka samman ämnesord som på olika sätt är besläktade på annat vis än hierarkiskt görs se även-hänvisningar, till exempel från ”Ornitologi” till ”Fåglar” och från ”Brottsutredning” till ”Detektiver”. Förklarande anmärkningar förtydligar vid behov hur en term ska användas, vilket ”kan gälla vaga ord som när de används som ämnesord har en mer entydig betydelse” (s. 85). Exemplet som gives lyder: ”Läromedelsgranskning. Anmärkning: Hit endast myndighetskontroll av läromedel. Andra undersökningar av läromedel förs till Läromedel.”

I samma riktlinjedokument (Kungliga biblioteket 2017a:10–18) ges också anvisningar för själva indexeringsarbetet. ”[A]lla typer av verk, t.ex. facklitteratur, artiklar, skönlitteratur, barn- och ungdomslitteratur, bilder” kan indexeras med Svenska ämnesord, men man medger samtidigt att ”[d]et är rimligt att ge ’tyngre’ publikationer en viss prioritet”, utan att närmare specificera vilka (s. 12). Som riktmärke för uttömmandegraden används regeln att ett ämnesord sätts bara om minst 20 procent av verket behandlar ämnet. Det finns i övrigt ingen begränsning för hur många ämnesord som får användas per dokument. De ämnesord som väljs ska vara de mest specifika möjliga, vilket till exempel innebär att boken *How children learn to read* indexeras med ”Läsinlärning”, inte ”Läsning”. Samma dokument indexeras inte med både över- och underordnad term ur samma hierarki – en bok om algebra får ämnesordet ”Algebra” men inte därtill ”Matematik”. Sådan indexering, förklaras det, hade gjort allmänna termer oanvändbara som sökord eftersom de också innefattade alla underordnade begrepp, och verk om allmänna ämnen hade blivit svårfunna.

Både pre- och postkoordinering används i Svenska ämnesord. I detta avseende har en viss förskjutning skett sedan systemets första inrättande, från prekoordinering mot postkoordinering. Nauri och Svanberg (2004:10) beskriver Svenska ämnesord som ett kort och gott prekoordinerat system, medan det enligt de nuvarande riktlinjerna (Kungliga biblioteket 2017a:12) är ”ett system som bygger på både prekoordinering och postkoordinering av ämnesord”. Systemets struktur baserades som nämnts på LCSH, ett i hög grad prekoordinerat system. Prekoordineringen i LCSH tar sig uttryck i en omfattande uppsättning underindelningar ordnade i olika kategorier: underindelningar knutna till ett visst ämnesord, som ”Helicopters – Flight testing”; geografiska underindelningar, som ”Tourism – California – San Francisco”; kronologiska underindelningar, som ”Engraving – 18th century”; formbestämningar, som ”Medicine – Periodicals”; och så kallade ”free-floating” underindelningar, en cirka 60 sidor lång lista över termer som med vissa begränsningar kan kombineras med valfritt huvudord (Library of Congress 2013a,b, 2015). Ett huvudord kan samtidigt ta underindelningar ur flera kategorier, till exempel ”United States – Foreign relations – 1783-1815 – Sources – Bibliography” (Library of Congress 2013b:5). Svenska ämnesord skapades enligt samma mönster.

Med tiden kom riktlinjerna för konstruktionen av ämnesordssträngar och användningen av underindelningar att uppfattas av både bibliotekarier och användare som alltför komplicerade, och med de nuvarande riktlinjerna från 2012 förenklades systemet genom införandet av en fasetterad struktur inspirerad av ämnesordssystemet FAST (Faceted Application of Subject Terminology).[7] Sålunda särskils åtta olika fasetter – allmänna

[7] Viktoria Lundborg, personlig kommunikation 2017-02-13.

ämnesord, personer, institutioner, händelser med formellt namn, titlar, tid, platser och genre/form – som var och en motsvarar ett särskilt datafält i katalogposten. Ett dokument som enligt de tidigare reglerna skulle indexeras ”Broderier – Osmanska riket – konferensmaterial” i samma fält indexeras nu med ”Broderier” i fältet för allmänt ämnesord, ”Osmanska riket” i fältet för geografiskt ämnesord, och ”Konferensmaterial” i fältet för genre/form-term. Underindelningar används fortfarande, men bara under allmänna och geografiska ämnesord, och underindelningen måste tillhöra samma fasett som huvudordet (Kungliga biblioteket 2017a:19). För allmänna underindelningar finns en särskild lista över godkända termer. I alla fasetter utom de allmänna och kronologiska samt genre/form används naturligt egennamn som ämnesord. Dessa finns inte i ämnesordlistan, utan kontrolleras mot en särskild auktoritetsdatabas eller konstrueras vid behov efter speciella riktlinjer för namnformer (Kungliga biblioteket 2017a:20). Godkända kronologiska ämnesord och genre/form-termer finns i särskilda listor. Utöver underindelningar finns prekoordinerade inslag i ämnesordlistan i form av sammansatta termer som ”Individen och samhället” (Kungliga biblioteket 2017a:79f).

# 3 Tidigare forskning

Här ska först diskuteras tidigare forskning inom uppsatsens ämnesområde, och därefter den tidigare forskningen och litteraturen om Svenska ämnesord.

## 3.1 Tidigare forskning om indexering

I äldre tiders kortkataloger var ämnesordssökning av nödvändighet en aktiv sökstrategi: den som sökte information om fotboll kunde välja att leta på ämnesordet ”Fotboll” och hitta alla dokument som indexerats med termen. Fungerade inte det kunde användaren prova med närliggande termer ur vokabulären. Man kunde också försöka att exempelvis bläddra bland titlar som börjar med ”Fotboll”, men inte samtidigt återvinna dokument som behandlar ämnet utan att skriva ut det i titeln. Samma begränsningar finns inte i dagens digitala bibliotekskataloger, där användaren enkelt kan söka på nyckelordet ”fotboll” och därmed återvinna alla dokument som har ordet som ämnesord, i titeln eller någon annanstans i katalogposten. Det är också vad användare har vant sig vid att göra; en rapport från OCLC (2009:12) konstaterar att ”[e]nd users want to be able to do a simple Google-like search and get results that exactly match what they expect to find”.

När i princip vad som helst i en katalogpost kan fungera som SAP är det naturligt att olika idéer framförs om vilka som är de bästa teknikerna för ämnesåtkomst. Detta avspeglas i en debatt för och emot kontrollerade vokabulärer, vilken sammanfattas väl av Gross, Taylor och Joudrey (2015). Ämnesorden betraktas av vissa som överflödiga i förhållande till annan metadata, och indexering med kontrollerade vokabulärer anses vara en föråldrad teknik. Förslag till alternativ är att utnyttja taggning utförd av användarna själva (Macgregor och McCulloch 2006), att automatiskt lägga till innehållsförteckningar, sammanfattningar eller annan metadata som kan bidra med ord för nyckelordssökning (Calhoun 2006), eller att kort och gott slopa ämnesorden i förvissningen att modern teknik gör dem obsoleta (Marcum 2005).

Debatten är förvisso inte ny, även om sentida tekniska landvinningar har fört den in på delvis nya vägar, utan kan spåras åtminstone till 1960-talet. Rowley (1994) delar in debatten om kontrollerade vokabulärer i dittills fyra epoker. Den andra epoken, och den egentliga debatten, inleddes i och med datorernas intåg och fick fart av de inflytelserika Cranfieldexperimenten (vars resultat sammanställdes av Cleverdon och Keen 1966). En mängd dokument indexerades med termer från 33 olika vokabulärer med varierande grad av kontroll och komplexitet, för att sedan testköras mot en samling sökfrågor, varigenom de olika vokabulärernas förmåga att återvinna relevanta dokument jämfördes med hjälp av mått på recall och precision. Experimenten syntes visa att okontrollerade vokabulärer kunde ge minst lika bra resultat som kontrollerade, och följdes av ett flertal liknande studier med samma slutsats. Om denna epok menar Rowley (1994:112) att forskningen syftade till att ta reda på vilken typ av indexering som var bäst, snarare än att utforska hur kontrollerade och okontrollerade vokabulärer kunde komplettera varandra. Detta nyanseras under hennes tredje epok, som domineras av mindre omfattande och mer specifika studier, vilka i regel kom fram till att någon kombination av kontrollerade och okontrollerade termer gav bäst resultat. Byrne (1975) är ett exempel: en jämförelse av testsökningar i en viss vetenskaplig databas på titlar, abstracts och ämnesord i olika kombinationer ledde författaren till slutsatsen att ”when dealing with engineering literature, titles by themselves do not convey enough information to allow the searcher to make a positive identification of a pertinent reference”, varför ”a reasonable compromise would be to use only titles and subject headings with free-language enhancers” (s. 229). Forskningen under Rowleys (1994) fjärde och samtida epok plockade upp dessa resultat och började göra studier på riktiga användare, vilket blivit särskilt aktuellt i och med den snabba utvecklingen av OPAC:er, gränssnitt, hyperlänkar, CD-ROM och andra tekniska nymodigheter.

Sedan dess har ett antal studier gjorts som belyser ämnesordsindexeringens värde och funktion genom att undersöka kontrollerade indextermers förhållande till annan metadata, till exempel titlar och innehållsförteckningar. Voorbij (1998) gjorde en tudelad studie med syftet att jämföra den relativa potentialen hos sökning med hjälp av ämnesord och sökning med hjälp av nyckelord ur dokumenttitlar. I den första delstudien fick tolv ämnesbibliotekarier inom humaniora och samhällsvetenskap välja ut vardera fyrtio katalogposter inom sitt ämne, och ombads för varje post bedöma överensstämmelsen mellan ämnesord och ord i titeln, uttryckt med hjälp av en sjugradig skala. Det skildes mellan vad som med Svenska ämnesords terminologi skulle kallas allmänna ämnesord, geografiska ämnesord och genre/form-termer. Poängen 4–7, motsvarande en låg överensstämmelse mellan ämnesord och titel, gällde för 40 % av de allmänna ämnesorden, 52 % av de geografiska och 78 % av genre/form-termerna. En följande bedömning av ämnesordens faktiska förbättring av katalogposterna gav att 37 % av posterna var ”considerably enhanced” och 49 % ”slightly or considerably enhanced”, också detta enligt ämnesbibliotekariernas uppfattning. I den andra delstudien fick samma försökspersoner göra sökningar inom sina ämnesområden, först med hjälp av nyckelord i titlarna, sedan med hjälp av ämnesord. Antalet under en timme återvunna dokument som bedömdes relevanta noterades. Av alla återvunna relevanta dokument hittades 86,9 % genom ämnesordssökningarna och 48,2 % genom titelnyckelordssökningarna. Vid sökning på breda ämnen var ämnesordssökningarnas relativa förtjänst gentemot titelsökningarna ännu större (87,8 % mot 41,0 %), och för smala ämnen något mindre (85,8 % mot 57,4 %).

Xu och Lancaster (1998) undersökte hur många SAP:er som tillförts av att en digital katalog gjort titlar och klassifikationskoder sökbara utöver ämnesorden. Härvid användes en utförlig regelsamling för att avgöra vilka SAP:er som skulle anses synonyma (exempelvis deweykod 598 och ”Birds”, eller ”stamp collecting” och ”philately”). 205 katalogposter studerades, varur 844 unika SAP:er identifierades. Cirka 75 % av dessa fanns i ämnesord, 54 % i titlar och 48 % i klassifikationskoder. Av de SAP:er som återfanns i bara en av fälttyperna bidrog ämnesorden med cirka 50 %, titlarna med 29 % och klassifikationskoderna med 21 %. Författarna drog slutsatsen att ”titles add only modestly to subject headings alone, and classification numbers contribute very few access points not provided by the other fields” (s. 66).

Nowick och Mering (2003) jämförde 3 275 fritextsökningar på en webbsida inriktad på vattenkvalitet med LCSH och två kontrollerade vattenvokabulärer. Mellan 30 och 40 % av sökningarna hade en exakt matchande term i någon av vokabulärerna, en siffra som rörde sig uppemot 60 % när små variationer i stavning och ordformer ignorerades. De visade också att cirka 85 % av sökningarna på denna vattenwebb gjordes på allmänna ämnen, 8 % på geografiska termer, 3–4 % på genre/form-termer och resterande på titlar, personnamn och annat.

Ansari (2005) undersökte 506 persiska doktorsavhandlingar i medicin, närmare bestämt inom ämnena pediatrik, gynekologi, kardiologi och psykiatri. Hon jämförde ämnesorden de tilldelats med deras titlar, och fann att 34,2 % av ämnesorden återfanns ordagrant i titeln, 5,7 % delvis och 30,5 % genom en se-hänvisning. Psykiatriavhandlingarna stack ut med sina 50 % exakta matchningar mellan ämnesord och titel, vilket enligt författarens spekulation vore ”probably due to use of more common words in Psychiatry titles” (s. 413).

Zavalina (2014) studerade metadata som beskriver samlingar av digitaliserat kulturarvsmaterial vid tre digitala bibliotek. Hon jämförde fritextbeskrivningar med ämnesord i 99 katalogposter. Graden av ”komplementaritet” mellan fritext och ämnesord mättes, termen åsyftande när det ena fältet innehåller information som inte uttrycks i det andra. Bland ämnesorden skildes också mellan allmänna, geografiska och kronologiska, samt genre/form-termer (återigen med Svenska ämnesords terminologi). Komplementariteten mättes åt båda hållen: fritext som kompletterar ämnesord och ämnesord som kompletterar fritext. Fritextbeskrivningarna visade sig komplettera allmänna ämnesord i 83 % av posterna, genre/form-termer i 49 %, kronologiska ämnesord i 46 % och geografiska i 29 %. Omvänt kompletterades fritextbeskrivningarna av allmänna ämnesord i bara 52 % av posterna, och av genre/form-termer i 23 %; de kronologiska ämnesorden kompletterade fritexten i 43 % av posterna, liksom de geografiska, vilka sistnämnda alltså var ensamma om att komplettera fritexten mer än denna kompletterade dem. I 40 % av posterna förekom komplementaritet mellan fritext och ämnesord åt båda hållen samtidigt. Att fritexten i snitt kompletterade ämnesorden mer än omvänt förklarades med fritextbeskrivningarnas längd, och med de höga graderna av komplementaritet ansåg författaren sig visa att ”more detailed collection-level metadata records which include both free-text and controlled-vocabulary subject metadata allow a fuller representation of the intellectual content of information objects and ultimately improve subject access for the users” (s. 87).

Gross, Taylor och Joudrey (2015) gjorde testsökningar i ett universitetsbiblioteks katalog med hjälp av 227 söktermer hämtade ur en logg, för att undersöka hur stor andel av de återvunna posterna som hade uteblivit utan ämnesorden närvarande. Ämnesorden antogs vara nödvändiga för en posts återvinning om minst ett ord ur söksträngen återfanns endast i ett ämnesordsfält (till exempel ”horror” ur sökningen ”horror films”). Resultaten visade att i snitt 27 % av sökträffarna hade uteblivit utan ämnesord. Siffran sjönk något till 24,8 % när sökträffarna begränsades till engelskspråkiga dokument. Samtliga poster i undersökningen hade berikats med innehållsförteckning och -sammanfattning. I en tidigare studie utförd i samma katalog innan berikningen hade Gross och Taylor (2005) kommit fram till att 35,9 % av träffarna fallit bort om inte för ämnesorden; fritexttilläggen sänkte alltså andelen från en dryg tredjedel till omkring en fjärdedel.

Maurer och Shakeri (2016) studerade indexeringen av elektroniska uppsatser och avhandlingar vid ett universitet, där de jämförde humaniora, samhällsvetenskap och de så kallade STEM-disciplinerna (naturvetenskap, teknik, ingenjörsvetenskap och matematik).[8] Både fria indextermer valda av författarna och katalogisatörernas ämnesord (ur LCSH) undersöktes. I båda fallen gällde att dokument inom humaniora i regel hade flest termer per post, STEM-disciplinerna minst, och samhällsvetenskap hamnade däremellan. Humaniora visade en större bredd i användningen av olika typer av ämnesord (personnamn, geografiska ämnesord med mera); inom både samhällsvetenskap och STEM-disciplinerna förekom nästan bara allmänna ämnesord. ”These findings are not surprising given the nature of the disciplines and the nature of LCSH”, konstaterade författarna, ”but they are interesting to document.” (s. 234)

Eftersom de refererade studierna använder olika metodologi och undersöker ämnesordsindexeringen från skilda perspektiv är det svårt att göra några direkta jämförelser mellan deras rent kvantitativa resultat, men de visar genomgående att ämnesorden gör nytta för användaren vid sökning. Ämnesordens för- och nackdelar gentemot fri indexering eller beskrivning har ett flertal författare bemödat sig om att sammanfatta, däribland Petras (2006) och Rowley (1994). Fördelar som ofta nämns med fritext är den lägre kostnaden, den högre potentiella specificiteten och uttömmandegraden, lättheten att välja termer, ingen väntan på att nya termer ska läggas till vokabulären och inga kompatibilitetsproblem mellan olika system. Fördelar med ämnesord är till exempel att synonymer och variantstavningar förs samman, homografer särskils, hänvisningar görs mellan besläktade termer och att kontrollen av valet av termer skapar högre precision. Generellt skulle alltså kunna sägas att frihet ger flexibilitet, medan kontroll ger struktur.

## 3.2 Publicerat om Svenska ämnesord

Svenska ämnesord är som tidigare konstaterats inte välbeforskat. Den tryckta litteraturen om systemet begränsar sig huvudsakligen till några kortare skrifter av handbokstyp. Nauri och Svanberg (2004) beskriver Svenska ämnesords (dåvarande) uppbyggnad och funktion, och även kortfattat dess historia, vilket också den under framtagandet av ämnesordlistan publicerade handledningen av Hellsten och Rosfelt (1997) gör. En lite personligare berättelse på samma ämne ges av Berg och Leth (2006). Utvecklingen av databasen och vokabulären lämnade efter sig ett antal numera svåråtkomliga projekt-

[8] Det blir faktiskt ”STEM” på engelska: *science, technology, engineering and mathematics*.

planer, protokoll och dylikt. KB tillhandahåller via sin hemsida en kort beskrivning av dagens Svenska ämnesord, samt riktlinjer och utbildningsmaterial (Kungliga biblioteket 2017a,b, u.å.).

Det största forskningsarbete som berör Svenska ämnesord torde vara den av Samuelsson (2008) författade avhandling i vilken systemets förmåga att uttrycka feministiska perspektiv undersöks (den är dålig). Liknande analyser görs i några examensarbeten av Sundin (2004), Folkesson och März (2006) och Nääs (2012). Det är även i övrigt till magister- och masteruppsatser man får vända sig för att hitta studier. Esperk (2001) jämför Svenska ämnesord i dess ännu inte helt färdiga form med några system som används i Storbritannien, och Lárusdóttir (2003) gör en jämförelse med LCSH och en tesaurus för konst och arkitektur. Tomic (2008) kommer i sin studie av bland annat Svenska ämnesords ”IR-effektivitet” fram till att dess styrka är förbättringen av recall, men att systemet kan förbättras genom tillägg av ordförklaringar, större syntaktisk och administrativ flexibilitet, och genom att länkas samman med fler vokabulärer än LCSH. Svenska ämnesord nämns gärna ytligt i diskussioner om ämnesordlistan för skönlitteratur, så till exempel i Aagaard och Viktorssons (2014) historiska redogörelse för denna; härom finns också ett antal examensarbeten vilkas uppräkning jag avstår från.

# 4 Teori

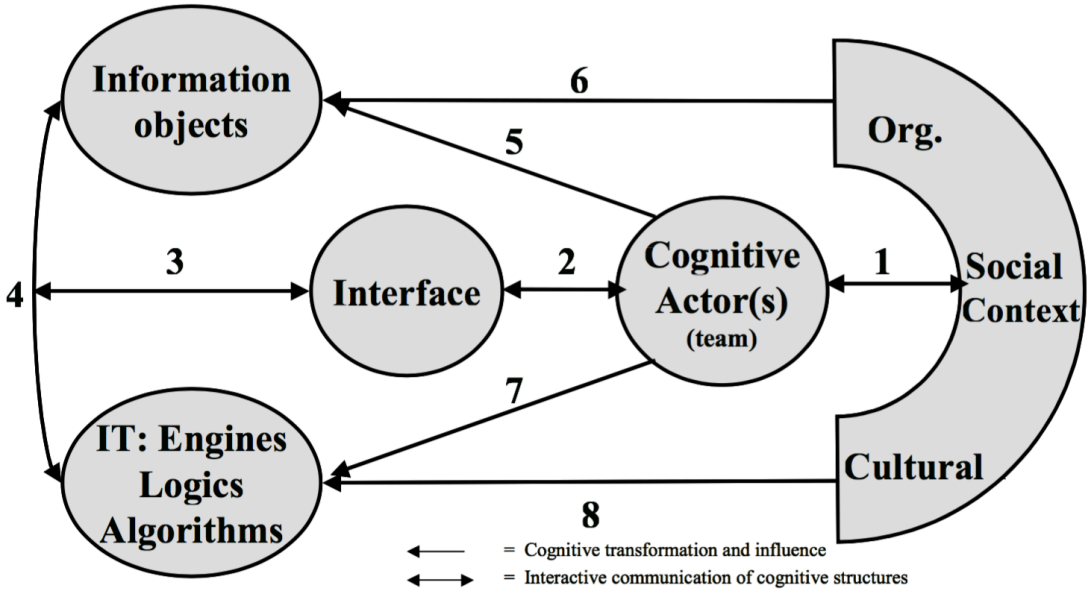
Med hjälp av ett ramverk som förklarar forskningsfältet placeras studien i ett akademiskt sammanhang. Därefter presenteras en modell för indelning av ämnesord i typer, vilken tillämpas i presentationen och analysen av undersökningens resultat.

## 4.1 Ramverk

Ämnesord och ämnesindexering kan studeras från olika perspektiv. Två vittfamnande forskningsområden av relevans är kunskapsorganisation (KO) och informationsåtervinning (IR, *information retrieval*). Ett KO-perspektiv har i någon mån skisserats i bakgrundsavsnittet. Ett ämnesordssystem ses som en särskild typ av bibliografiskt språk, vars uppgift är att beskriva och systematiskt ordna dokument med hjälp av syntaktiska och semantiska regler (se vidare Svenonius 2000).

Med ett IR-perspektiv sätts återvinningen av dokument i centrum. Ingwersen och Järvelin (2005) identifierar två huvudsakliga paradigm inom IR-fältet. Den traditionella IR-forskningen är tekniskt inriktad, med fokus på att statistiskt mäta sökalgoritmers och andra teknologiers förmåga att återvinna dokument. Studierna är i regel testkörningar av algoritmer på en samling dokument, och algoritmerna jämförs utifrån återvinningsgraden av dokument som på förhand bestämts vara relevanta. Denna ”laboratoriemodell” har sedan 1990-talet kritiserats för bristande realism, i det att ingen hänsyn tas till användarnas faktiska sökbeteenden och interaktion med systemen. En forskningsgren har därmed uppstått som istället fokuserar på sökprocessen i ett vidare sammanhang, där användarens sökstrategier och informationskontext hamnar i centrum.

Ingwersen och Järvelin (2005) presenterar en schematisk bild av IR-forskningens studieobjekt, vars mål är att integrera de olika synsätten (se figur 1). Med hjälp av detta




Figur 1: IR-forskningens beståndsdelar enligt Ingwersen och Järvelin (2005:261).

schema kan en studie tydligare kontextualiseras. Figuren illustrerar hur olika delar av en informationssökningsstruktur förhåller sig till varandra och interagerar. ”Cognitive actors”[9] inrymmer en mängd roller: författare, indexerare, systemtekniker, gränssnittsdesigners, informationssökare med flera – alla individer eller grupper av individer med någon relevans för sökprocessen. Samma person kan inta olika roller vid olika tillfällen, och skifta roll med korta intervaller. Dessa så kallade kognitiva aktörer framträder ur och interagerar med sin respektive sociala kontext (pil 1), som beroende på individens aktuella roll kan bestå av dennes kollegor eller av målgruppen för det system som studeras. ”Information objects” är alla de ting som eftersöks vid informationssökning, i princip vad jag kallat ”dokument”. Dessa skapas av kognitiva aktörer (5) under inflytande av deras kontext (6); de bearbetas, organiseras och återvinns (4) av de tekniska systemen (”IT: engines, logics, algorithms”), vilka förses med gränssnitt (”Interface”) med vilka människor interagerar (2). Teknikens behandling av dokumenten är till stor del beroende av mänsklig katalogisering, indexering med mera (5).

I den här uppsatsen kan katalogposterna ses som informationsobjekt och Svenska ämnesord som IT, ett system som lägger till metadata i informationsobjekten. Studien undersöker därmed modellens interaktion 4 – vilken nytta ämnesorden tillför informationsobjekten – med sökbarhet i fokus. Modellens övriga roller och interaktioner är relevanta, men behandlas inte direkt.

## 4.2 Analysdimensioner

Studiens syfte är att jämföra olika ämnesord och diskutera skillnaderna i hur de överlappar med titel och innehållsförteckning.[10] För att det ska vara meningsfullt måste äm-

[9] Innebörden av begreppet ”kognitiv” i sammanhanget kan förbigås utan allvarliga konsekvenser.

[10] Titel och innehållsförteckning kallar jag här med en gemensam beteckning för *fritext*, en term som kan ha en vidare innebörd i andra sammanhang.

nesorden delas in i olika grupper, som kan jämföras med varandra. Eftersom tidigare forskning i regel undvikit att jämföra olika typer av ämnesord saknas en etablerad modell för indelningen. Jag föreslår därför en modell enligt vilken en mängd ämnesord kan delas in i grupper på fyra olika sätt, och analyseras längs respektive dimensioner: en *funktionell*, *ämnesmässig*, *formell* respektive *abstraktionsnivå*dimension.

Den funktionella dimensionen innebär att ämnesorden grupperas efter den typ av information de innehåller, så att till exempel allmänt ämne, tid, plats och genre/form skiljs från varandra. I katalogformatet MARC motsvarar dessa grupper var sin typ av fält i katalogposten (de olika 6XX-fälten: 600, 610, 611 osv.), vilket gör uppdelningen tekniskt okomplicerad. Underindelningar i delfält kan också lätt skiljas från huvudord. En indelning av det här slaget görs av både Voorbij (1998) och Zavalina (2014), vilka utifrån denna diskuterar ämnesordens förhållande till fritext, och av Maurer och Shakeri (2016), som dock stannar vid att redovisa antalet ämnesord inom varje grupp.

En ämnesmässig gruppering innebär att varje ämnesord förs till ett övergripande ämne, så att exempelvis ”Kaktusar” och ”Näbbmöss” förs till biologi, medan ”Determinism” och ”Gettierproblemet” förs till filosofi. Vilka grupper som är relevanta kommer att skilja sig mellan olika kataloger, urval och beroende på studiens syfte. Förmodligen är indelningen bara möjlig att meningsfullt göra med ämnesord som anger ett allmänt ämne (till exempel kan kronologiska ämnesord som ”1900-talet” och ”Medeltiden” knappast föras till olika ämnesområden), och även där öppnar det för mycket godtycke. Ansari (2005) gör en typ av ämnesmässig indelning när hon jämför ämnesord med titlar inom fyra olika medicinska delområden, men baserar den inte på ämnesorden som sådana utan på dokumentens disciplinära tillhörighet.

En formell indelning av ämnesorden innebär att de grupperas till exempel efter grammatisk form (singular kontra plural, bestämd kontra obestämd form), längd eller antal ord (så som ”Ledarskap” är ett ord men ”Personlig utveckling” två).

Abstraktionsnivå syftar på ämnesordens placering i en tänkt hierarki från abstrakt till konkret, till exempel ”Djur”–”Hästar”–”Ardenner”. Den inbyggda hierarkiska strukturen i vokabulären kan här vara till hjälp. I de flesta fall torde det dock inte vara så lätt som att helt enkelt räkna ämnesordens avstånd från topp- eller bottennivån, eftersom den hierarkiska detaljrikedomen kan variera mellan olika delar av systemet. Det är också i allmänhet svårt och kanske meningslöst att jämföra abstraktionsnivån mellan vitt skilda företeelser så som ”Framgång”, ”Andligt sökande” och ”Navigeringsutrustning”.

Fler slags indelningar är garanterat möjliga, och vissa mer användbara än andra. De ovan föreslagna kan kombineras så att till exempel en analys längs den ämnesmässiga dimensionen görs endast hos en viss funktionell kategori, eller så att en formell indelning sker bara på en viss abstraktionsnivå. Poängen är att ämnesorden i en katalog inte är en homogen massa, och att detta åskådliggörs genom att man ur olika perspektiv skiljer dem från varandra.

# 5 Metod

För varje katalogpost har ämnesorden jämförts med titel och innehållsförteckning, och överlappningen har förts in i en tabell. Olika typer av ämnesord jämförs med utgångspunkt i deras varierande uppträdande i denna statistiska sammanställning. Pearsons $chi^2$-test har använts för att testa samband, och i ett fall en linjär regressionsanalys. Som gräns för signifikans används den gängse nivån 5 % (Eggeby och Söderberg 1999). Testen har gjorts i statistikprogrammet SPSS.

## 5.1 Avgränsningar och urval

De katalogposter som studeras hämtades ur den svenska nationalbibliografin, med hjälp och medgivande från KB:s katalogsupport. Nationalbibliografin är en databas där KB registrerar alla böcker som ges ut i Sverige, och därtill tidskrifter, kartor, rapporter med mera. Det icke förlagsutgivna som insamlas har en viss slagsida åt humaniora och samhällsvetenskap (Kungliga biblioteket 2015:4). Detta gäller ju även för Svenska ämnesord. Valet av nationalbibliografin som studieobjekt beror på katalogiseringens höga kvalitet och noggrannhet. I en föregående pilotstudie med bredare omfång påträffades återkommande brister i katalogiseringens konsekvens och överensstämmelse med riktlinjer, vilket inte har varit fallet i det urval som använts för den här studien. Valet innebär samtidigt att den ämnesindexering som studeras inte är representativ för indexeringen vid andra svenska bibliotek än det kungliga.

Inom nationalbibliografin avgränsades urvalet vidare till:

- tryckt facklitteratur,
- på svenska,
- katalogiserad mellan 2012 och 2016,
- med innehållsförteckning och minst ett ämnesord.

Avgränsningen till facklitteratur gjordes i enlighet med Svenska ämnesords huvudsakliga inriktning. Skön- och facklitterär ämnesindexering har ganska olika förutsättningar, och att samtidigt undersöka båda vore ett komplicerat projekt. Den skönlitterära ämnesordlistan är ännu inte heller fullt integrerad med Svenska ämnesord, även om en sådan process har inletts (Kungliga biblioteket 2017c).

Att se endast till svenskspråkig litteratur är naturligt eftersom det är ämnesordens språk. Att svenskspråkiga ämnesord ökar sökbarheten för en bok med titel på engelska eller tyska är tämligen uppenbart. Voorbij (1998:468) gör en motsvarande avgränsning till nederländska titlar, men motiverar det med spekulationen att exempelvis franska titlar vore ”inclined to be more vague”; det gör jag inga anspråk på att uttala mig om.

Den bortre gränsen vid 2012 sattes för att de nuvarande riktlinjerna för Svenska ämnesord infördes detta år. Den indexering som gjorts med de gamla riktlinjerna är inte omedelbart jämförbar med den senare. Bland de studerade posterna upptäcktes likväl en handfull som följde den äldre standarden; dessa korrigerades manuellt.

Nationalbibliografins svenska facklitteratur katalogiserad mellan 2012 och 2016 uppgår till drygt 30 000 poster.[11] Den nödvändiga avgränsningen till poster med både ämnesord och innehållsförteckning minskade antalet till cirka 1 500, eller en tjugondel. Det innebär att en förhållandevis liten delmängd av samlingen faktiskt studeras. Med tanke på den låga förekomsten av innehållsförteckningar är det vanskligt att dra alltför vidlyftiga slutsatser om ämnesordens nytta eller onödighet utifrån graden av överlappning med dessa.

En vidare filtrering reducerade de 1 500 posterna till 1 348, när de poster som innehöll ämnesord men inte Svenska ämnesord sorterats bort; dessa decimerades ytterligare till 1 283 när några katalogiseringsdatum äldre än 2012 beklagligen uppdagades. Urvalsstorleken bestämdes med hjälp av en standardtabell för statistiskt representativa urval (Krejcie och Morgan 1970:608) till knappt 300, vilket för enkelhets skull avrundades till 300 exakt. Posterna valdes slumpmässigt, men proceduren komplicerades genom att urvalet utan egentlig anledning stratifierades efter katalogiseringsår. En ursprunglig idé att också undersöka fältet för innehållssammanfattning hade lett till att vissa av de hämtade posterna innehöll en sådan, men saknade innehållsförteckning. 27 sådana poster fanns bland de 300 utvalda, och ersattes med 27 andra slumpvalda poster med innehållsförteckning. Det slumpvalet gjordes utan stratifiering.

## 5.2 Analys

Analysen av ämnesorden i katalogposterna vilar på deras indelning i grupper, och jämförelserna mellan ämnesord och fritextfält måste ske systematiskt.

### 5.2.1 Tillämpning av analysdimensioner

Den tidigare presenterade modellen för indelning av ämnesorden i grupper har anpassats till det aktuella urvalet. Den funktionella dimensionen motsvaras här av de fasetter som används i Svenska ämnesord: allmänna ämnesord, personnamn, institutionsnamn, händelser med formellt namn, titlar, tid, plats och genre/form; därtill också distinktionen mellan huvudord och underindelning. Det är den enda analysdimension som kunnat tillämpas utan inskränkningar. En formell indelning har vidare gjorts bland de allmänna ämnesorden, som grupperats efter längd räknad i antal tecken samt antal ord. Detta ansågs inte meningsfullt för de andra kategorierna, till exempel de geografiska ämnesorden, där skillnaden i längd mellan ”Håbo” och ”Kristinehamn” inte torde innebära någon skillnad i sannolikheten att ordet förekommer i titel eller innehållsförteckning. Bland de geografiska ämnesorden kunde däremot en sorts abstraktionsnivåuppdelning göras, av typen ”Sverige”–”Halland”–”Laholm”. Ämnesmässig uppdelning visade sig praktiskt ogenomförbart.

Tabell 1 visar med ett fiktivt exempel de olika ämnesordsfasetter som används i Svenska ämnesord, motsvarande uppdelningen efter funktion. Under det allmänna, det geografiska och titelämnesordet syns användning av underindelningar. Underindelningar av allmänna ämnesord (”allmänna underindelningar”) finns i Svenska ämnesord i en särskild lista om ett fyrtiotal termer, och detsamma gäller de kronologiska ämnesorden och

[11] Sökningen *(nb2012mon OR nb2013mon OR nb2014mon OR nb2015mon OR nb2016mon) spr:swe STIL:0 NOT mat:eresurs* i Libris (http://libris.kb.se) ger 33 272 träffar [2017-03-31].

*Hårda tag : biskopens attityd till mässbesökare*

| | |
|---|---|
| personnamn | Söderblom, Nathan, 1866–1931 |
| institutionsnamn | Svenska kyrkan |
| händelse | Baltiska utställningen |
| titel | Bibeln – G.T. – Domarboken |
| kronologiskt | 1910-talet |
| allmänt | Brottsbekämpning – religiösa aspekter |
| geografiskt | Sverige – Skåne – Malmö |
| genre/form | Lättläst |

Tabell 1: Exempelanvändning av Svenska ämnesords fasetter.

genre/form-termerna, vilka dock är något fler. För personnamn, institutionsnamn, händelser med formellt namn, titlar och geografiska ämnesord finns av naturliga skäl inga direkta begränsningar. En särskild auktoritetsdatabas styr i vissa fall valet av namnform (Kungliga biblioteket 2017a).

### 5.2.2 Redundansmätning

För varje ämnesord i en katalogpost har noterats om det förekommer också i titeln eller i innehållsförteckningen. Eftersom undersökningen avser ämnesordens förbättring av sökbarheten genom tillförandet av SAP:er räknas ett ämnesord inte som redundant utifrån sitt informationsinnehåll, utan utifrån sin form. Detta skiljer sig från tidigare utförda studier, till exempel Xu och Lancaster (1998) som räknar ämnesordet ”USA” som upprepat i titeln om denna nämner ”America” och menar samma sak. Jag har inte räknat ämnesordet ”Talsymbolik” som redundant för att titeln eller innehållsförteckningen innehåller ”numerologi”, eftersom en sökning på det ena ordet inte ger träffar på det andra.

Även om redundansen bestäms utifrån en ren jämförelse av textsträngar finns det oklara fall och mellanting. Vissa bibliotekskataloger kan till exempel föra samman grammatiska former så att ”pyssel” och ”pysslet” ger samma sökträffar. Ämnesord som består av flera ord, till exempel ”Fornegyptisk religion”, kan ha det ena ordet i titel eller innehållsförteckning men ändå tillföra en SAP med det andra. För att göra analysen mer flexibel har jag skilt mellan fullständig och ofullständig redundans, där fullständig redundans innebär:

- Ett ämnesord bestående av ett ord förekommer som ett fristående ord i exakt samma form i fritext. Parentestillägg ignoreras.

  | | |
  |---|---|
  | Filantroper | *Att ge: samtal med svenska filantroper* |
  | Val (psykologi) | *Medvetna val : från offerkofta till möjlighetsmantel* |

- Ett ämnesord bestående av flera ord har båda orden någonstans i samma fritextfält. Småord som ”och” och prepositioner ignoreras.

| | |
|---|---|
| Biologisk mångfald | ”[…] biologisk mångfald […]” |
| Barn och musik | ”[…] Musik för alla barn? […]” |

- Ett personnamn som ämnesord förekommer med för- och efternamn i ett fritextfält, med eller utan särskiljande levnadsår.

| | |
|---|---|
| Tegnér, Alice, 1864-1943 | *Alice Tegnér : musikskapare* |

Ofullständig redundans innebär:

- Ett ämnesord bestående av ett ord förekommer som en del av ett annat ord, eller i en närliggande grammatisk form.

| | |
|---|---|
| Fotboll | ”[…] samtliga sluttabeller i seriefotbollen genom åren […]” |
| Visioner | *Bön, vision, skapande : i Herrens vingård* |

- Ett ämnesord bestående av flera ord har åtminstone ett ord som ensamt skulle räknas som ofullständigt redundant, eller har något fullständigt redundant ord men inte alla.

| | |
|---|---|
| Ekonomisk politik | ”[…] ekonomerna och den ekonomiska teorin […]” |
| Medicinsk mikrobiologi | ”[…] Klinisk mikrobiologi […]” |

Förekomsten av ämnesord i titel respektive innehållsförteckning har noterats separat. Redundansen både hos unika ämnesord och bland den totala mängden ämnesord har räknats.

Bedömningarna har gjorts manuellt för varje katalogpost. En tänkbar möjlighet vore att automatisera processen med hjälp av någon form av algoritm som kunde matcha ämnesord mot fritext, och säkerligen producera mängder av fängslande data. Det skulle tillåta att ett betydligt större urval undersöktes, eller i princip att en hel katalog genomlöptes maskinellt. Utvecklingen av ett sådant datorprogram ligger dock bortom såväl min egen kompetens som uppsatsarbetets tidsomfång. För att sätta upp och finjustera maskinläsbara regler för redundansmätningen skulle dessutom krävas god förtrogenhet med materialet, en förtrogenhet som åtminstone i mitt fall har vuxit fram under det manuella granskningsarbetet.

# 6 Resultat och analys

De 300 undersökta katalogposterna innehåller sammanlagt 1101 ämnesord, eller 742 unika (när varje ämnesord räknas bara en gång oavsett antalet förekomster). Det innebär att varje post har i genomsnitt 3,67 ämnesord, och att varje unikt ämnesord används 1,48 gånger.

| | förekomster | poster | äo/post | unika äo | anv./äo |
|---|---|---|---|---|---|
| allmänt (huvudord) | 627 | 291 | 2,2 | 546 | 1,2 |
| – underindelning | 91 | 80 | 1,1 | 18 | 5,1 |
| geografiskt | 185 | 116 | 1,6 | 61 | 3,0 |
| genre/form | 78 | 70 | 1,1 | 28 | 2,8 |
| kronologiskt | 49 | 32 | 1,5 | 21 | 2,3 |
| personnamn | 45 | 37 | 1,2 | 45 | 1,0 |
| institutionsnamn | 16 | 14 | 1,1 | 15 | 1,1 |
| titel | 10 | 5 | 2,0 | 8 | 1,3 |
| händelse | 0 | 0 | 0 | 0 | 0 |

Tabell 2: Översikt över ämnesord efter funktion.

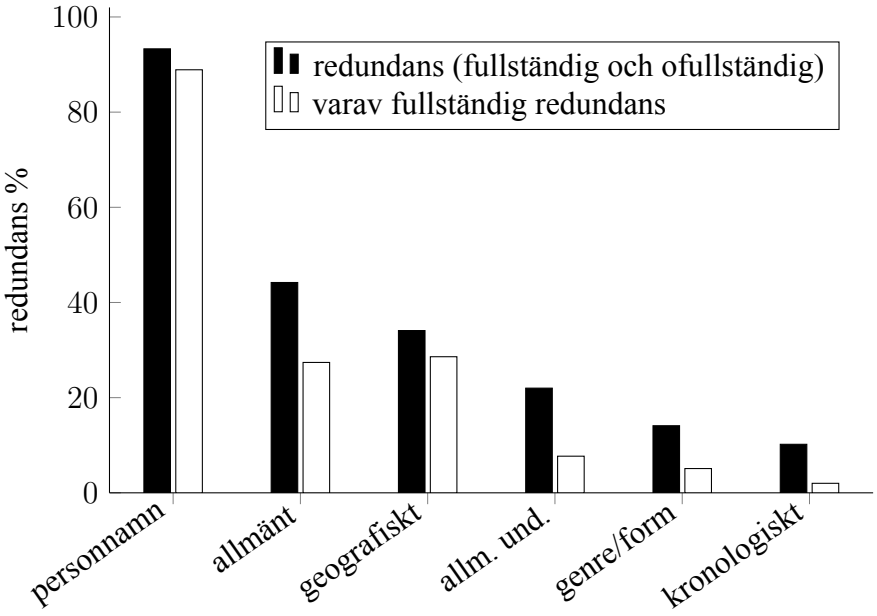
## 6.1 Skillnader efter funktion

I tabell 2 redovisas ämnesorden uppdelade efter funktion: hur många gånger de förekommer i urvalet, hur många poster de uppträder i, hur många ämnesord som i snitt förekommer per post, hur många unika ämnesord som finns per kategori, och hur många gånger varje unikt ämnesord genomsnittligen förekommer i hela urvalet. Underindelningar till allmänna ämnesord redovisas separat från huvudord, så även i det följande; övriga underindelningar är här sammanräknade med huvudorden så att exempelvis ”Sverige – Närke – Örebro” räknas som tre geografiska ämnesord. Institutionsnamn, titel och händelse är sällsynta och utelämnas från vidare analys.

Figur 2 visar för varje kategori andelen redundanta förekomster av ämnesord. Vid sidan av andelen ämnesord som uppvisar redundans överhuvudtaget, fullständig eller ofullständig, redovisas andelen ämnesord som i något av fritextfälten uppvisar fullständig redundans.

Personnamn har med god marginal högst redundans. Av 45 förekomster som ämnesord finns 42 i titel eller innehållsförteckning (93,3 %), varav 40 fullständigt. Det är inte så förvånande. Knappt hälften av verken med personnamnsämnesord är biografier (av genre/form-termen ”Biografi” att döma), och bland de resterande finns till exempel utställningskataloger, fotoböcker och brevsamlingar. Det är dessa typer av publikationer som handlar tillräckligt mycket om en person för att motivera namnet som ämnesord, och de tenderar också att i titeln ange personens namn. Till skillnad från allmänna ämnesord finns få möjligheter att använda olika synonymer i fritext och ämnesord, eller välja mellan närliggande termer – Marie Göranzon heter just det och inget annat. Det förklarar också att nästan all personnamnsredundans är fullständig (undantagen är Dante Alighieri, som ju gärna kallas bara vid förnamn, och Kurt Hamrin, som heter Kurre i innehållsförteckningen).

I den andra änden finns de kronologiska ämnesorden, med bara 5 redundanta förekomster av 49 (10,2 %). 15 av de 49 är ”1900-talet”, och av de 21 unika ämnesorden är 13 ett sekel eller årtionde enligt samma format. Så ser de flesta godkända kronologiska termerna ut i Svenska ämnesord, men det är väl ingen långsökt tanke att historisk litte-




Figur 2: Redundans efter funktion.

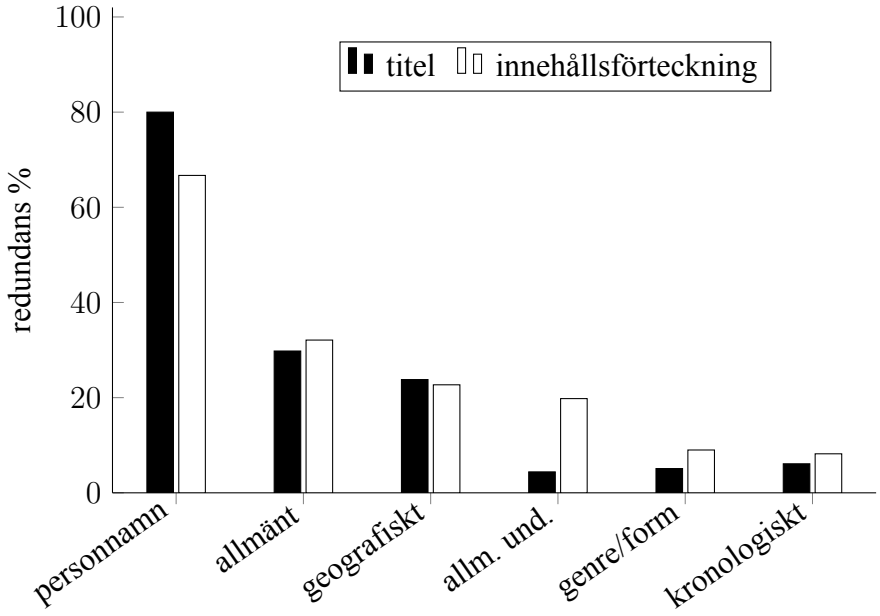
ratur hellre använder sig av mer exakta periodiseringar som är relevanta för det aktuella verket. Riktlinjerna tillåter som sista utväg fritt formulerade årtalsspann som ämnesord, och alla tre sådana i urvalet uppvisar redundans (katalogisatören blev nog frestad av innehållsförteckningens tydliga tidsangivelse).

Genre/form-termerna uppvisar också en låg redundans: 11 av 78 förekomster är redundanta (14,1 %). 21 av de 78 är ”Biografi”, och inte ens detta ord förekommer mer än en gång i en titel och en gång i en innehållsförteckning. ”Handledning” och ”Intervjuer” förekommer någon gång i en titel, men för det mesta är det nog ovanligt och onödigt att en bok själv anger vilken genre den tillhör. Det vore till exempel överflödigt att sätta undertiteln ”läromedel” till *Gymnasiekemi 2*. Genre/form-termerna är dessutom ofta i pluralform (”Kokböcker”, ”Utställningskataloger”), så någon fullständig redundans blir det ännu mer ogärna.

De allmänna underindelningarna förekommer 91 gånger i urvalet, och är 20 gånger redundanta (22,0 %). 40 av 91 förekomster är ”historia”. Av 18 unika ämnesord uppvisar fem alls någon redundans: ”historia”, ”juridik och lagstiftning”, ”sociala aspekter”, ”teori, filosofi” och ”framställning, tillverkning etc.”. Om dessa termer har något gemensamt som gör dem mer benägna att användas i fritexten än till exempel ”forskning”, ”psykologiska aspekter” eller ”metodik” är det inget jag lyckas identifiera, och det kan förstås vara en slump. Att den fullständiga redundansen är låg i förhållande till den totala torde förklaras av ämnesordens form – till exempel det återkommande ”aspekter”, ett ord som i de flesta andra sammanhang för en lite mer anonym tillvaro.

Mellan de tre grupperna med lägst redundans finns inbördes ingen statistiskt signifikant skillnad[12]. Det är dock intressant att notera att dessa tre typer av dokumentbeskriv-

[12] $p = 0{,}156085 > 0{,}05$, $df = 2$, $n = 218$.




Figur 3: Redundans efter funktion, uppdelad på upprepande fält.

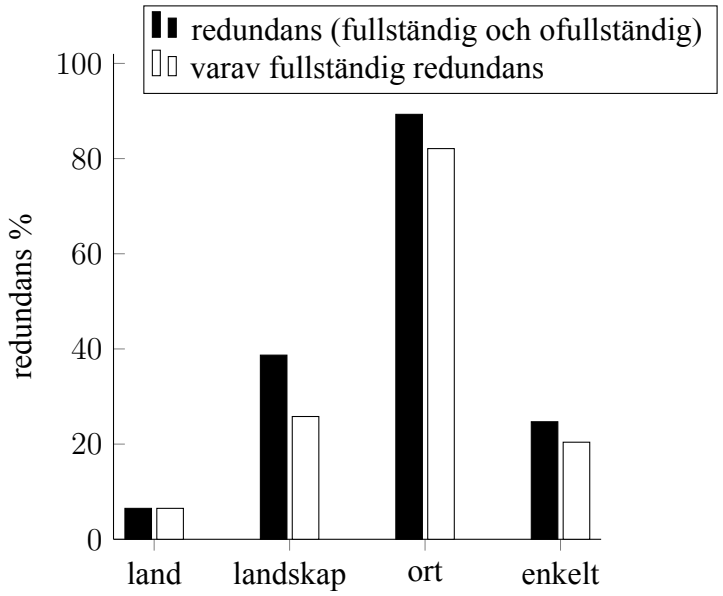
ningar – allmänna underindelningar, genre/form-termer och kronologiska ämnesord – förhåller sig mer abstrakt till verket självt än vad åtminstone personnamnen och de allmänna ämnesorden gör. Personnamnen beskriver med maximal exakthet en specifik individ som behandlas i texten, och de allmänna ämnesorden anger begrepp som visserligen kan vara ganska övergripande (”Ekonomi”, ”Medicin”), men ofta mer precisa (”Fornegyptisk religion”, ”Lusthus”), och som i vilket fall som helst ligger nära verkets självuppfattning. De kronologiska ämnesorden approximerar den egentliga tidsbestämningen med en uppsättning standardperioder, genre/form-termerna beskriver innehållet bara indirekt via generella kategorier, och de allmänna underindelningarna är en typ av metaämnesord som i första hand beskriver ett annat ämnesord. Att det intellektuella avståndet mellan ämnesord och verkets sakinnehåll korrelerar med det terminologiska är ganska lätt att acceptera. En annan viktig iakttagelse är att de tillåtna termerna i de tre kategorierna med lägst redundans är betydligt färre än de allmänna ämnesorden, och som personnamn kan vilket namn som helst anges. Ju färre termer som finns att välja på, desto sämre är möjligheten att pricka in samma ordval som i verket självt.

De geografiska ämnesorden är inbördes heterogena och behandlas därför separat nedan.

Figur 3 visar andelen förekomster av ämnesord i titel respektive innehållsförteckning. Endast för de allmänna underindelningarna är skillnaden signifikant.[13] Redundansen är låg i bägge fallen, men uppenbart förekommer den typ av vaga och svepande ord som syns i underindelningarna hellre i innehållsförteckningar än i titlar, som gärna försöker vara lite roligare än så. Ett exempel är *Jämlik ålderdom? : samtiden och framtiden*, med kapitelrubriken ”Den sociala konstruktionen av ojämlikt åldrande”, där underindelningen ”sociala aspekter” delvis återfinns.

[13] $p = 0{,}001456 < 0{,}05$, $df = 1$, $n = 91$.

## 6.2 Skillnader efter geografisk nivå




Figur 4: De geografiska ämnesordens redundans efter geografisk nivå.

Användningen av geografiska ämnesord ser i materialet ut på två olika sätt. Enkel indikation av ett land eller annan större geografisk enhet förekommer 93 gånger, varav 75 förekomster är ”Sverige”. För snävare platsangivelser kräver riktlinjerna för Svenska ämnesord att de högre nivåerna också inkluderas. För orter inom Sverige gäller att de måste föregås av ämnesordet ”Sverige” plus landskap: ”Sverige – Värmland – Kristinehamn”; utomlands räcker land plus ort. 31 gånger förekommer ett land i sådant sammanhang, och 30 gånger är det ”Sverige”. Nationalbibliografins geografiska intresseområde är rätt uppenbart.

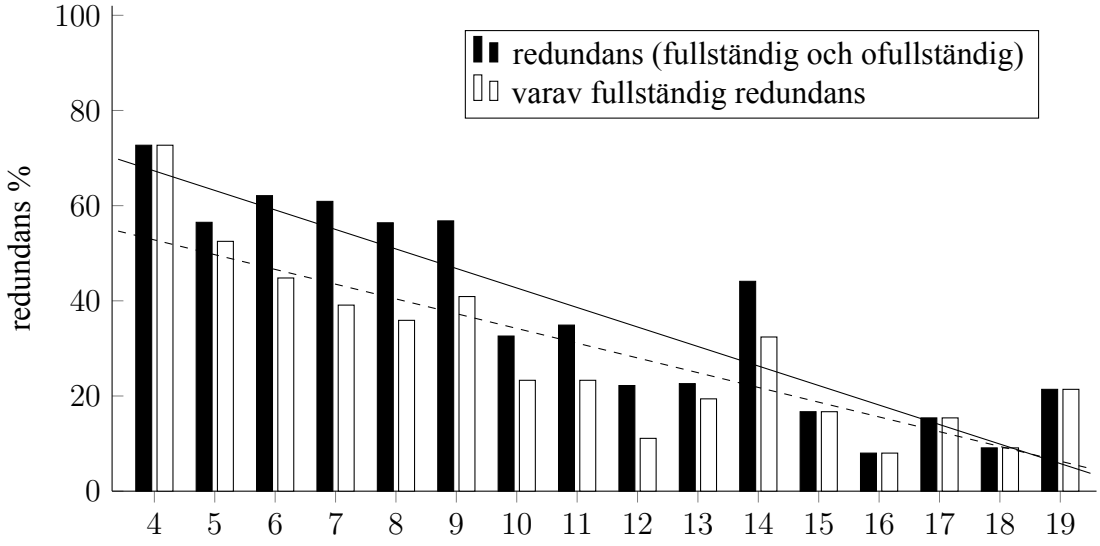
Skillnaden i redundans för de tre nivåerna redovisas i figur 4. Landet är redundant bara i 6,5 % av fallen, landskapet i 38,7 %, och orten i hela 89,3 %. Det är närmast övertydligt hur redundansen ökar med den geografiska specificiteten. Det förklaras också enkelt. Ett typiskt exempel är *Morden i Glimminge 1886 : i dokument och pekoral*: att Glimminge ligger i Skåne vore ett onödigt påpekande för den som är intresserad av ämnet (och än mer att Skåne ligger i Sverige), men helt enligt regelverket har boken ändå indexerats ”Sverige – Skåne – Glimminge”. Det torde gälla generellt att ett verk på svenska om en särskild svensk ort anser sig handla i första hand om den orten, i vissa fall också om det omgivande landskapet (till exempel kokboken *Från raggarkorv till älgfilé : Värmland, Torsby, människorna, maten, skogen*), och i princip aldrig om landet Sverige.

Bakom stapeln för de enkla geografiska ämnesorden döljer sig 20,0 % redundans för ”Sverige”, vilket dock inte motsvarar någon statistiskt signifikant skillnad gentemot ”Sverige” som första term i en ortsangivelse.[14] Också för verk som behandlar svenska företeelser och förhållanden utan särskilt ortsfokus gäller att landet för det mesta är en

[14] $p = 0{,}093825 > 0{,}05$, $df = 1$, $n = 105$.

outtalad bakgrund (till exempel *Statsministerns sommarläsning : om litteratur, politik och bildning*, om svenska politikers läsvanor). Utländska platser som enkla geografiska ämnesord är redundanta till 44,4 %, vilket *är* en signifikant skillnad mot ”Sverige”:s 20,0 %[15], helt i linje med tanken att det i svenskutgivna böcker är mer relevant att i titel eller innehållsförteckning nämna vilket land man talar om när det inte är vårt hemland.

## 6.3 Skillnader efter längd




Figur 5: De allmänna ämnesordens redundans efter längd.

I figur 5 visas redundansen för de allmänna ämnesorden, uppdelad efter ämnesordets längd i tecken (”Etik” är fyra tecken långt och ”Dagstidningar” tretton). Endast enordiga ämnesord har tagits med, och parentestillägg har räknats bort (”Brott mot män” finns inte medräknat, och ”Bön (kristendom)” har räknats som tre tecken); ordantalen analyseras nämligen för sig nedan, och parentestilläggen har ju ignorerats vid redundansbedömningen. Figuren visar den totala förekomsten av ämnesord, men värdena för unika ämnesord är snarlika. Sällsynta ordlängder ger osäkra redundansvärden, och därför visas bara ämnesordslängder med minst tio förekomster.

Det syns en tendens för redundansen att falla med ämnesordens ökande längd, och en regressionsanalys visar att den är signifikant.[16] Tendensen kan spekulativt förklaras. Korta ord är ofta enkla begrepp med allmän betydelse som inte gärna kan eller behöver formuleras i omskrivningar – de vanligaste fyra, fem och sex tecken långa ämnesorden är ”Kemi”, ”Döden” och ”Gruvor”. Långa ord är däremot ofta sammansättningar med en mer komplex innebörd som tillåter större formuleringsvariation i fritexten – de vanligaste 17, 18 och 19 tecken långa ämnesorden är ”Alkoholkonsumtion”, ”Katastrofberedskap” och ”Anställningsvillkor”. Å andra sidan är dessa termer samtidigt exempel på tekniska begrepp med tydliga definitioner, och det är lätt att föreställa sig en ovilja

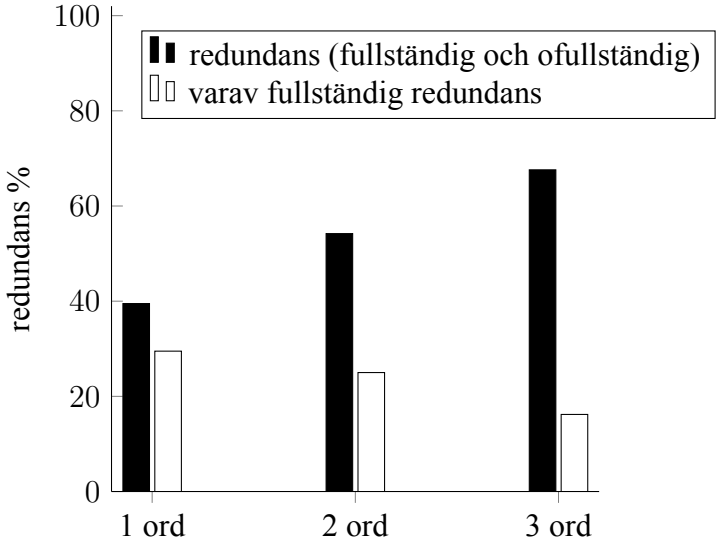
[15] $p = 0{,}03088 < 0{,}05$, $df = 1$, $n = 93$.

[16] $p = 0{,}000002 < 0{,}05$, $r^2 = 0{,}815194$ för redundans överhuvudtaget; $p = 0{,}000041 < 0{,}05$, $r^2 = 0{,}710931$ för fullständig redundans i något av fritextfälten.

att ersätta facktermer med omformuleringar i mer seriöst betonad litteratur. Tolkningen är inte uppenbar.

Den iögonenfallande höga stapeln för de fjorton tecken långa ämnesorden kan jag inte förklara med något annat än slumpen.

## 6.4 Skillnader efter ordantal




Figur 6: De allmänna ämnesordens redundans efter antal ord.

De flesta ämnesord består av ett enda ord, men också två ord (”Palliativ vård”), tre ord (”Offentliga elektroniska tjänster”) eller flera (”Personer med psykisk funktionsnedsättning”) förekommer. Fler än tre ord är dock väldigt sällsynt förekommande i urvalet och har inte tagits med i figur 6, som visar redundansen för ett, två och tre ord långa allmänna ämnesord. Liksom i tidigare figurer visas de totala förekomsterna av ämnesord (värdena för unika ämnesord är genomgående några procentenheter högre).

Synbarligen ökar sannolikheten att ett ämnesord är redundant ju fler ord det består av, men sannolikheten sjunker samtidigt att det är fullständigt redundant. Det verkar intuitivt rimligt: ”Amerikansk experimentell poesi” har ju tre möjligheter att åtminstone delvis överlappa med något ord i titeln eller innehållsförteckningen, men för fullständig redundans krävs å andra sidan att alla tre nämns. Ett statistiskt test visar emellertid att skillnaden mellan de höga två- och treordsstaplarna inte är signifikant, och sambandet mellan ordantal och fullständig redundans är inte signifkant överhuvudtaget.[17] Att dessutom småord har ignorerats vid redundansmätningen försvårar ytterligare tolkningen av resultatet, då de treordiga ämnesorden nästan uteslutande består av två ord sammanlänkade med preposition eller ”och” (till exempel ”Barn med adhd” eller ”Mat och dryck”).

[17] För skillnaden mellan den sammanlagda redundansen hos två- och treordiga ämnesord gäller att p = 0,149629 > 0,05, df = 1, n = 128. För sambandet mellan ordantal och fullständig redundans gäller att p = 0,684556 > 0,05, df = 2, n = 537.

| hög redundans | låg redundans |
|---|---|
| personnamn | allmänna underindelningar |
| | genre/form-termer |
| | kronologiska ämnesord |
| korta ämnesord | långa ämnesord |
| orter | länder |

Tabell 3: Sammanfattning av redundanssambanden.

# 7 Slutsatser och diskussion

Inledningsvis presenterades fyra frågor som uppsatsen skulle besvara:

- Vilka typer av ämnesord tenderar att återfinnas också i titeln eller innehållsförteckningen?
- Vilka typer av ämnesord tenderar att *inte* återfinnas i titeln eller innehållsförteckningen, och därmed skapa unika SAP:er?
- Vilka är orsakerna till de funna mönstren i ämnesordens redundans?
- Vad betyder ämnesordens redundansmönster för deras nytta?

De första tre har vidrörts i det föregående avsnittet, men svaren förtjänar att sammanfattas.

”Typer av ämnesord” har förståtts på några olika sätt: funktion (allmänna ämnesord, geografiska, personnamn, …), längd i tecken, längd i ord, och för de geografiska ämnesorden deras geografiska nivå. Den funktionella kategori som uppvisade överlägset högst redundans gentemot både titel och innehållsförteckning var personnamnen, och lägst redundans sågs hos de allmänna underindelningarna, genre/form-termerna och de kronologiska ämnesorden. Korta allmänna ämnesord, mätt i tecken, var i högre grad redundanta än långa. Mätt i ord antyddes ett delvis omvänt förhållande, som dock inte bekräftades av ett statistiskt signifikanstest. De geografiska ämnesordens redundans var högre ju mindre område de beskrev. Tabell 3 sammanställer dessa resultat och besvarar därmed de två första frågorna ovan.

Dessa resultat är i sig själva intressanta. De blir förstås än mer intressanta av att förklaras. Det påpekades tidigare att de tre funktionella kategorierna med lägst redundans förhåller sig förmedlat till verkets ämnen: via beskrivningen av ett annat ämnesord, av verkets form snarare än innehåll, och av en generell tidsperiod. De är på så vis abstrakta typer av beskrivningar, medan allmänna ämnesord och inte minst personnamn står i direkt förbindelse med innehållet i verket. Distinktionen mellan högredundanta orter och lågredundanta länder kan ses som en motsvarande skillnad mellan konkret och abstrakt

hos de geografiska ämnesorden. Det är frestande att i detta läsa in ett generellt samband mellan abstraktion och låg redundans, som skulle innebära att de mer specifika ämnesorden inom varje funktionell kategori uppvisade en högre redundans än de mer allmänt hållna, så att ”Toypudlar” och ”1960-talet” vore oftare redundanta än ”Hundar” och ”1900-talet”. Detta har inte varit praktiskt genomförbart att undersöka i den här studien, men vidare forskning skulle kunna ge tydligare svar. Redundansförhållandet mellan korta och långa ämnesord stämmer dåligt överens med denna hypotes, i att långa ämnesord vanligen är mer specifika än korta, men ändå visar en lägre redundans.

Den låga redundansen hos allmänna underindelningar, genre/form-termer och kronologiska ämnesord hänger som tidigare konstaterats troligen ihop med att de tillgängliga termerna inom dessa kategorier är färre än för de allmänna ämnesorden och personnamnen, vilket minskar sannolikheten att indexeraren hittar samma ordval som titeln eller innehållsförteckningen. Detsamma kan inte förklara mönstret i de geografiska ämnesordens redundans, där ingen begränsning finns i antalet tillgängliga termer. Här är istället Svenska ämnesords riktlinjer för geografisk ämnesindexering en bättre förklaring, som ju kräver att ort anges tillsammans med land. Den högre redundansen hos korta ämnesord i förhållande till långa har för sin del förklarats med att längre ord lättare kan ersättas med omformuleringar och omskrivningar, men detta är spekulation och skulle behöva bekräftas av en mer djupgående undersökning. Något enkelt svar på frågan om orsakerna till de identifierade redundansmönstren verkar inte kunna ges. I vilken mån mönstren kan generaliseras till ämnesordsindexering utanför studiens avgränsningar kan bara vidare forskning ge svar på.

Resultaten är förenliga med tidigare forskning, i den mån de är jämförbara. Voorbij (1998) studerade allmänna ämnesord, geografiska ämnesord och genre/form-termer, och fann samma inbördes förhållande mellan redundansen hos dessa tre som i min studie. Han mätte den emellertid annorlunda, och dristade sig inte till att förklara skillnaderna i redundans. Zavalina (2014) jämförde samma ämnesordstyper, och därtill kronologiska, men undersökte komplementaritet istället för redundans och hennes resultat är svåra att jämföra med den här uppsatsens.

Som påpekades i uppsatsens inledning ska ”redundant” inte tolkas som ”onödig”, och hög redundans hos en viss grupp ämnesord är inte i sig ett kvitto på deras onytta. Samma rapport från OCLC (2009:12) som konstaterar att användare vill kunna söka fritt som på Google beskriver också hur de uppskattar möjligheter till mer avancerade sökningar, med stöd av verktyg som fasettering av sökträffar. Många bibliotekskataloger länkar ämnesorden till en sökning på termen i katalogen, och för på så vis samman dokument om samma ämne. Det är möjligt att olika typer av ämnesord tjänar olika primära syften. Person- och ortnamnens höga redundans gör att deras nytta som extra SAP vid nyckelordssökning är begränsad, men funktionen att länka samman relevant litteratur kvarstår. Svenska ämnesords riktlinjer och relaterade dokument för ingen särskild diskussion om ämnesordens olika syften och användningsområden, vilket kanske inte är nödvändigt om ämnesord självklart tas för ett avsiktligen begagnat IR-system. Om ämnesorden istället snarare fungerar som komplement till nyckelordssökning uppstår behovet att mer detaljerat utreda hur de används. Vidare forskning behövs för att utröna om den faktiska användningen av exempelvis länkning och fasettering med ämnesord i bibliotekskataloger skiljer sig mellan olika typer av ämnesord. Detta i kombination med

vetskapen om deras relativa redundans kan bidra till mer ändamålsenlig design av katalogernas gränssnitt och funktioner, och även skapa förutsättningar för mer grundade överväganden och prioriteringar vid utvecklingen av befintliga och framtida ämnesordlistor.

Uppsatsen har lämnat ett teoretiskt bidrag till den önskade framtida forskningen om ämnesordstypers skilda egenskaper genom att föreslå några olika så kallade dimensioner längs vilka ämnesord kan analyseras, och tillämpa dessa. Det vore intressant att se en fortsatt utveckling av idén. En studie av större format kunde gärna ta sig an att jämföra ämnesord med fritext på algoritimisk väg.

Frånsett detaljerna i ämnesordstypernas redundans gentemot fritexten visar resultaten att ämnesorden tagna som helhet tillför katalogposterna ett mervärde: ökad sökbarhet. Redundansen är under 50 % för alla termer utom person- och ortnamn, vilket innebär att mer än hälften av dessa ämnesord har skapat en SAP som annars inte funnes. Detta ensamt försvarar inte bruket av ämnesordsindexering, utan hänsyn måste också tas till arbetsinsatsen och de resurser indexeringen tar i anspråk – överväganden som bara kan göras från fall till fall.

# Tack

Jag vill först och främst tacka min handledare Jonas Fransson för hans hjälp under hela arbetets gång. Kursansvarige Olof Sundin förtjänar också ett tack. Därtill skänker jag ett stort tack till personalen vid samhällsvetenskapliga institutionens bibliotek vid Lunds universitet, där jag under min praktikperiod inspirerades till denna studie. Slutligen vill jag tacka Libris katalogsupport, som bistått med de katalogposter som studerats, och Lars Wahlgren, som gett värdefull statistisk hjälp.

# Referenser

Aagaard, Harriet & Viktorsson, Elisabet (2014). Subject headings for fiction in Sweden: a cooperative development. *Cataloging & Classification Quarterly*, 52(1), s. 62–68.

Ansari, Mariam (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review*, 54(7), s. 410–414.

Badke, William (2012). Save the subject heading. *Online*, 36(6), s. 48–50.

Berg, Ingrid & Leth, Pia (2006). Ämnesordsindexering i Sverige: hur ett nationellt system baserat på internationella standarder skapades i ett land som ofta velat gå sina egna vägar. I af Malmborg, Kerstin (red.) – *Jag heter Lena och jag har försökt göra något: barn, bibliotek och visioner: en vänbok till Lena Lundgren till 60-årsdagen den 19 mars 2006*. Stockholm: Kulturförvaltningen, s. 167–172.

Byrne, Jerry R. (1975). Relative effectiveness of titles, abstracts, and subject headings for machine retrieval from the COMPENDEX services. *Journal of the American Society for Information Science*, 26(4), s. 223–229.

Calhoun, Karen (2006). *The changing nature of the catalog and its integration with other discovery tools* (Rapport). Washington, D.C.: Library of Congress.

Cleverdon, Cyril & Keen, Michael (1966). *Factors determining the performance of indexing languages*. Cranfield: College of Aeronautics.

Eggeby, Eva & Söderberg, Johan (1999). *Kvantitativa metoder: för samhällsvetare och humanister*. Lund: Studentlitteratur.

Esperk, Evi (2001). *Verbal ämnesindexering i nationalbibliotekskataloger: en jämförelse mellan brittisk och svensk praxis*. Magisteruppsats, Högskolan i Borås.

Folkesson, Isabel & März, Klara (2006). *Marginaliserad kunskap?: en kritisk studie av representationen av genusvetenskaplig litteratur i klassifikationssystem och ämnesordslistor*. Magisteruppsats, Högskolan i Borås.

Gartner, Richard (2016). *Metadata: shaping knowledge from antiquity to the semantic web*. Schweiz: Springer.

Gross, Tina & Taylor, Arlene G. (2005). What have we got to lose?: the effect of controlled vocabulary on keyword searching results. *College & Research Libraries*, 66(3), s. 212–230.

Gross, Tina, Taylor, Arlene G. & Joudrey, Daniel N. (2015). Still a lot to lose: the role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly*, 53(1), s. 1–39.

Hellsten, Unn & Rosfelt, Margareta (1997). *Ämnesordsindexering: en handledning*. Stockholm: Kungliga biblioteket.

Hjørland, Birger (1993). *Emnerepræsentation og informationssøgning: bidrag til en teori på kundskabsteoretisk grundlag*. Borås: Valfrid.

Hjørland, Birger & Kyllesbech Nielsen, Lykke (2001). Subject access points in electronic retrieval. *Annual Review of Information Science and Technology*, 35, s. 249–298.

IFLA. *Se* International Federation of Library Associations and Institutions.

Ingwersen, Peter & Järvelin, Kalervo (2005). *The turn: integration of information seeking and retrieval in context.* Dordrecht: Springer.

International Federation of Library Associations and Institutions (1998/2006). *Funktionella krav på bibliografiska poster: slutrapport*. Stockholm: Svensk biblioteksförening.

Internationella standardiseringsorganisationen (1985). *ISO 5963:1985 Documentation – Methods for examining documents, determining their subjects, and selecting indexing terms*. Genève: Internationella standardiseringsorganisationen.

ISO. *Se* Internationella standardiseringsorganisationen.

Krejcie & Morgan (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 30, s. 607–610.

Kungliga biblioteket (2015). *Utgivningspuls 2015: nationalbibliografin i siffror* (Rapport). Stockholm: Kungliga biblioteket.
http://www.kb.se/dokument/2015.pdf [2017-03-31]

Kungliga biblioteket (2017a). *Riktlinjer för indexering med Svenska ämnesord*. Stockholm: Kungliga biblioteket. http://www.kb.se/Dokument/Verktygsladan/Svenska ämnesord/Riktlinjer/Riktlinjer SAO.pdf [2017-02-15]

Kungliga biblioteket (2017b). *Om Svenska ämnesord*. http://www.kb.se/katalogisering/Svenska-amnesord/om/ [2017-02-15]

Kungliga biblioteket (2017c). *Indexering av skön- och barnlitteratur i Libris*. http://www.kb.se/katalogisering/Svenska-amnesord/Indexering-i-LIBRIS/ [2017-03-23]

Kungliga biblioteket (u.å.). *Utbildning i ämnesordsindexering*. http://www.kb.se/dokument/Verktygsladan/Svenska ämnesord/Presentationer/Utbildning i Svenska ämnesord elev.pdf [2017-02-15]

Lancaster, Frederick Wilfrid (1991). *Indexing and abstracting in theory and practice*. London: Library Association.

Lárusdóttir, Álfheidur (2003). *Indexeringsspråk, indexerare och användare: en jämförelse av Library of Congress Subject Headings, Svenska ämnesord och The Art & Architecture Thesaurus*. Magisteruppsats, Uppsala universitet.

Library of Congress (2013a). *Library of Congress Subject Heading Manual, H 830: Geographic Subdivision*. Washington, D.C.: Library of Congress. https://www.loc.gov/aba/publications/FreeSHM/H0830.pdf [2017-02-16]

Library of Congress (2013b). *Library of Congress Subject Heading Manual, H 1075: Subdivisions*. Washington, D.C.: Library of Congress. https://www.loc.gov/aba/publications/FreeSHM/H1075.pdf [2017-02-16]

Library of Congress (2015). *Library of Congress Subject Heading Manual, H 1095: Free-Floating Subdivisions*. Washington, D.C.: Library of Congress. https://www.loc.gov/aba/publications/FreeSHM/H1095.pdf [2017-02-16]

Macgregor, George & McCulloch, Emma (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5), s. 291–300.

Marcum, Deanna B. (2005). *The future of cataloging: adress to the Ebsco leadership seminar.* Boston, Massachusetts, 16 januari 2005.

Maurer, Margaret Beecher & Shakeri, Shadi (2016). Disciplinary differences: LCSH and keyword assignment for ETDs from different disciplines. *Cataloging & Classification Quarterly*, 54(4), s. 213–243.

Nauri, Miriam & Svanberg, Magdalena (2004). *Svenska ämnesord: en introduktion*. Stockholm: Kungliga biblioteket.

Nowick, Elaine A. & Mering, Margaret (2003). Comparisons between Internet users' free-text queries and controlled vocabularies: a case study in water quality. *Technical Services Quarterly*, 21(2), s. 15–32.

Nääs, Lina (2012). *Den Andre i hyllan och på webben: benämnandets makt i sociala taggar och ämnesord knuta till HBTQ-relaterad skönlitteratur*. Masteruppsats, Uppsala universitet.

OCLC (2009). *Online catalogs: what users and librarians want* (Rapport). Dublin, Ohio: Online Computer Library Center.

Olson, Hope A. & Boll, John J. (2001). *Subject analysis in online catalogs*. 2. uppl., Englewood, Colorado: Libraries Unlimited.

Petras, Vivien (2006). *Translating dialects in search: mapping between specialized languages of discourse and documentary languages*. Diss., University of California, Berkeley.

Rowley, Jennifer (1994). The controlled versus natural indexing languagues debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), s. 108–118.

Rowley, Jennifer & Hartley, Richard (2008). *Organizing knowledge: an introduction to managing access to information*. 4. uppl., Hampshire: Ashgate Publishing.

Samuelsson, Jenny (2008). *På väg från ingenstans: kritik och emancipation av kunskapsorganisation för feministisk forskning*. Diss., Umeå universitet.

Soergel, Dagobert (1985). *Organizing information: principles of data base and retrieval systems*. Orlando, Florida: Academic Press.

Soergel, Dagobert (1994). Indexing and retrieval performance: the logical evidence. *Journal of the American Society for Information Science*, 45(8), s. 589–599.

Sundin, Maria (2004). *Identitet och ordning: om kunskapssyn och kunskapsorganisation i en bibliotekskontext*. Magisteruppsats, Uppsala universitet.

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

Taylor, Arlene G. & Joudrey, Daniel N. (2009). *The organization of information*. 3. uppl., Westport, Connecticut: Libraries Unlimited.

Tomic, Taeda (2008). *Thesauri or ontologies? Or both?: a comparison between two kinds of subject heading systems with regard to their enhancement of effective information retrieval*. Masteruppsats, Uppsala universitet.

Voorbij, Henk J. (1998). Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation*, 54(4), s. 466–476.

Xu, Hong & Lancaster, Frederick Wilfrid (1998). Redundancy and uniqueness of subject access points in online catalogs. *Library Resources & Technical Services*, 42(1), s. 61–66.

Zavalina, Oksana L. (2014). Complementarity in subject metadata in large-scale digital libraries: a comparative analysis. *Cataloging & Classification Quarterly*, 52(1), s. 77–89.