Nedskrivning av Webben

en mediearkeologisk analys av Heritrix

Simon Alfredsson

Examensarbete (30 högskolepoäng) i arkivvetenskap för
masterexamen inom ABM-masterprogrammet vid Lunds universitet.
Handledare: Lars Ilshammar
År: 2014

Title
Writing Down the Web: a Media Archaeological Analysis of Heritrix

Abstract
There is no doubt that the the Web constitutes an important channel and platform for public discourse and cultural exchange, and that our times will be remembered by accessing the archives of the Web. This is also to say, that perception of the web – present and past – is shaped and governed by the media we and our machines perceive it with. So with what media do we archive it? Recording media is a largely unexplored factor in the preservation of the web and archival science in general, and the question requires rethinking.

This Master's thesis examines a prominent instrument used for archiving the web – the software Heritrix – as a *technical recording medium*. As a contrast and background to Heritrix, the electronic legal deposit law recently enacted in Sweden is written as another medium used to transmit content published online. My research question is, how does Heritrix work as a archiving medium?

Heritrix is shown to be *technical* in the sense that it inscribes, without human intervention and without any respect for human discourse, signals that reach it, much like how a photographic plate indiscriminately registers and inscribes light by chemical reaction. It is a *recording* medium in the sense that the transmission traverses across time as well as space. To enable the study of Heritrix as medium, I employ a media archaeological method inspired mainly by the media theory of Friedrich Kittler. By carefully delimiting my study Heritrix is as far as possible considered on its own terms, to build an diagrammatic description of the medium and the way it fits into our contemporary discourse network.

Four epistemological configurations and one time perspective create a full description, and the disparity between the electronic legal deposit law and Heritrix is shown to be massive, with Heritrix being a technical medium while the electronic legal deposit law remains an artistic medium.

Keywords
Web crawler, medium, archival science, web preservation, media archaeology

INNEHÅLLSFÖRTECKNING

Inledning

Denna masteruppsats behandlar Heritrix, ett datorprogram som automatiskt besöker och hämtar hem webbsidor, som medium. Inom arkivvetenskapen behandlas medium ofta som neutrala databärare, något som jag uppfattar som en stor svaghet. Heritrix är ett viktigt arkiveringsmedium för webben, och med min studie öppnar jag en diskussion om mediets betydelse för arkiv som jag tror är viktig.

I detta introducerande kapitel ska jag kortfattat skriva *bakgrunden*[1] till min studie, inklusive mitt studieobjekt och dess kontext. Sedan följer min *problemformulering* där jag först beskriver problemområdet mitt problem ingår i, följt av den formella problemformuleringen och de avgränsningar den innebär, samt argument för de val jag gjort. Slutligen *sammanfattar* jag mitt syfte och mina mål med uppsatsen.

Bakgrund

Sveriges lag om pliktexemplar kan spåras till år 1661, och har fram tills nyligen inriktat sig på tryckt material – exemplar av allt som trycks i Sverige har sedan dess (i princip) skickats till Kungliga Biblioteket (KB) och andra stora bibliotek i Sverige som bevarar det för framtida forskare och andra. Sedan Internets genomslag som publiceringsplattform har en ny lag om elektronisk pliktleverans (den lag som till slut blev SFS 2012:492, och som härefter kallas e-plikten eller lagen om e-plikt) blivit aktuell, då en stor del av allt som publiceras i Sverige numera endast görs tillgängligt på Internet. Lagen har varit omdebatterad och diskuterad länge innan den infördes.

Lagstiftningen innebär att allt material som publicerat elektroniskt och som anses vara ”färdigt” eller ”avslutat” (vilket utesluter exempelvis dataspel) ska skickas in som pliktexemplar till KB. Ambitionen när man författade lagen verkar enligt Lars Ilshammar ha varit att insamlingen av detta nya medium ska ske på samma sätt som tidigare i så stor utsträckning som möjligt, och har för detta kritiserats som baksynt (Ilshammar 2014:2-3).

E-plikten är fortfarande ny, och speciellt de tekniska lösningar som kan komma att användas för att i så stor mån som möjlig automatisera och förenkla leveranser, både för leverantörer och KB, är fortfarande i högsta grad under utveckling. Även mer teoretiska eller juridiska frågor om vad som utgör ett publicerat dokument online, hur

1 I början av varje kapitel, och emellanåt i introduktioner till delar av kapitel, beskriver jag dispositionen som följer kort och *kursiverar* de ord som senare används som rubriker. Det ska även nämnas att min stil i denna uppsatsen, med mycket listor, relativt fristående avsnitt i analysen och andra särdrag som kan upplevas som konstiga, är medvetna försök att förena *innehåll* med *form* eller *teori* med *praktik;* mer om detta i mitt teori-avsnitt.

detta ska levereras till KB och med vilken metadata, samt hur det ska bevaras, är snårigheter som KBs avdelning för pliktleveranser försöker navigera.

Automatisk insamling av den svenska webben

Väl innan det byråkratiska maskineriet kring e-plikten ens börjat rulla skedde dock insamling av webben på KB. 1996 var man på KB väl medveten om att mycket av det kulturarv som publicerades online sedermera gick förlorat på grund av att insamling ej skedde, varför man startade Kulturarw³-projektet (kw³). Alla svenska sidor[2] besöks och sparas automatiskt av programvaran Heritrix, en så kallad *spindel* som börjar med att hämta en uppsättning på förhand givna sidor (kallade *seeds*) specificerade genom deras *Uniform Resource Identifier*[3] (URI) och sedan hittar nya sidor att besöka genom att rekursivt[4] följa hyperlänkar till andra URIs i hämtade dokument. På så sätt kan man spara ner stora delar av den synliga ej lösenordsskyddade delen av Internet (eller snarare webben)[5] som är tillgänglig via *Hyper Text Transfer Protocol* (HTTP) – med många undantag av olika tekniska skäl[6].

Spindlar används främst för att indexera webben, för att göra den sökbar med hjälp av sökmotorer. Vad som särskiljer Heritrix är att den utvecklats av och för insamlande institutioner – programvaran är ett resultat av ett samarbete mellan The Internet Archive och diverse nordiska nationalbibliotek – och designad för att inte bara indexera utan arkivera webben på ett sätt som uppfyller krav kring långtidsbevarande. Programvaran är fritt licensierad[7] och kan köras på vilken dator som helst, även om större jobb (såsom att arkivera hela den svenska webben, eller hela webben) kräver

---

2 Vilka sidor som är svenska fastställts internt och arbiträrt inom projektet, kortfattat inkluderas alla webbplatser publicerade med de svenska domänsuffixen .se och .nu, samt webbplatser vars servrar finns inom det geografiska Sverige. Helst hade man velat använda mer omfattande kriterier såsom att samla in allt som ”rör svenska intressen”, men kriterierna måste anpassas efter tekniska förutsättningar (korrespondens med Allan Arvidsson på KB [Mottagen 5 dec 2013]).

3 URI är en kompakt unik sträng som används för att namnge en resurs på ett nätverk såsom Internet. *Uniform Resource Locator* (URL) är en subtyp av URI och är det vi oftast skriver in i adressfältet på vanliga webbläsare när vi vill nå en resurs, exempelvis http://www.dn.se/. Se https://en.wikipedia.org/w/index.php?title=Uniform_resource_identifier&oldid=594378170 [Hämtad 14 feb 2014]. För att säkerställa att mina läsare ser samma wikipedia-artikel som jag, länkar jag emellanåt till en specifik version av artikeln. Att använda wikipedia-artiklar i sådan usträckning som jag gör i min uppsats kan ses som kontroversiellt, men som Kirschenbaum säger så är Wikipedia en av de mest uppdaterade och bästa källorna när det kommer till tekniska ämnen, och därför är hänvisning dit befogat även i vetenskapliga publikationer (2008:xvii).

4 Rekursion innebär att något gör något med sig själv, i programmeringssammanhang kan det exempelvis handla om att en funktion anropar (startar) sig själv upprepade gånger. Ett annat exempel på rekursivt beteende är när ett program (såsom *find*, se https://www.gnu.org/software/findutils/) söker igenom hierarkier av mappar på ett filsystem genom att gå igenom varje mapps submappar, deras submappar etc., eller länkar på webbsidor, och sedan alla länkar på de sidorna etc. ”Om du inte förstår rekursion, läs den här meningen igen” är ett skämt som ska förklara rekursion, se https://sv.wikipedia.org/wiki/Rekursion.

5 Internet är en term som innefattar även andra protokoll förutom HTTP, exempelvis GOPHER och USENET. Webben använder jag som en mer precis term som innefattar endast HTTP (i andra sammanhang kan termen webben även inkludera exempelvis protokollen FTP samt RSS).

6 Till exempel har spindlar problem att på ett tillfredsställande sätt arkivera sidor som skapas dynamiskt ur databaser eller som använder JavaScript eller andra tekniker för att generera och presentera innehåll. Webbens enorma storlek, diversitet och snabba utveckling, och den stora mängd webbsidor som publiceras hela tiden, utgör andra påfallande problem som alla diskuteras flitigt inom datavetenskapen, se exempelvis Spaniol, Denev, Mazeika, Weikum, Senellart (2009). Spindelns relation till sin omgivning analyseras grundligt i mitt resultat och analys-avsnitt.

mer hårdvara än smalare insamling (såsom att arkivera alla sidor från en eller ett par domäner).

Medium i arkiven

Traditionellt är en viktig fråga för arkivarier att ta ställning till hanteringen och valet av medium i vilket handlingar ska lagras i eller migreras till (Östholm 1995:54-5). Medium måste utvärderas ur flera synvinklar, inklusive hur varje medievals tekniska beskaffenhet påverkar sökbarheten, hållbarheten, långtidsbevarandet, tillgängligheten till och säkerheten hos arkivet, och dessa praktiska aspekter av medier är forskningsområden som inom arkivvetenskap behandlas utförligt. Det är också främst sådana aspekter som Riksarkivet diskuterar i sina föreskrifter. Riksarkivet strävar alltjämt efter att behandla handlingar som frikopplade från medium och databärare (RA-FS 1991:1), vilket också är drömmen inom digitalt långtidsbevarande, se exempelvis Quisbert (2008) och Giaretta (2011). IT-system av alla sorter, programmering och programvaruutveckling överallt, och de flesta andra system vi rör oss i, vilar på en *objektorienterad* filosofi med modularitet och inkapsling[8] som honnörsord, något som jag återkommer till.

Även inom byråkratin är valet av medium för arkivhandlingar först och främst en *teknisk* fråga, där egenskaper relateras till det övergripande målen om bevarande, tillgänglighet och ordning som föreskrivs i lagen. Men finns där inte mer att säga om de medium och databärare som huserar och representerar våra arkiv? I och med den nya e-plikten kan man på KB ställa sig frågan, vad gör dom egentligen på kw³ som inte kommer att tas hand om med den nya lagen? Enligt Allan Arvidsson på kw³ har projektet länge efterfrågat utökade resurser, utan större intresse från ovan[9]. Hur fungerar Heritrix? Kan en djupare medieteori hjälpa oss till en bättre insikt?

Man kan argumentera för att det är innehållet i webbarkiven som är det väsentliga, men mot det kan man ställa Allan Arvidssons anekdot, återgivet från vår korrespondens den 9 januari 2014, inklusive några felstavningar,

> Vad forskare kommer att vara intressant kan vi bara gissa. Man kan här dra paralleller till medeltida handskrifter. När jag först kom till KB så blev jag lite förvånad. Det finns forskare som är intresserade av papperskvaliten, andra tittar tekninken för inbindningen etc. Det är inte alls säkert att det är någon som är intresserad av vad det står i boken! På samma sätt är det kanske med webbarkiv?

Min studie

Detta avsnitt börjar med att introducera den studie jag kommer bygga min uppsats kring genom att först ge problemformuleringen, följt av ett avsnitt med andledningar

7 Licensen heter Apache License, Version 2.0 (se http://www.apache.org/licenses/LICENSE-2.0), vissa delar av källkoden är licensierad under andra (fria) licenser. Man kan ladda ner programmet och dess källkod på https://webarchive.jira.com/wiki/display/Heritrix/Heritrix [Hämtad 14 feb 2014].

8 För introduktioner till object-orienterad programmering, inklusive diskussion om modularitet och encapsulation, se https://en.wikipedia.org/w/index.php?title=Object-oriented_programming&oldid=595599010 [Hämtad 19 feb 2014]. Se även mitt avsnitt om paradigmet i resultat och analys-avsnittet nedan.

9 Korrespondens via e-post mottagen den 9 jan 2014.

till *varför* det är ett angeläget problem att studera. Efter detta görs en översiktlig introduktion till min *data*, en serie *avgränsningar*, följt av ett övergripande syfte och andra *resultat* jag vill att min studie ska åstadkomma.

Problemformulering

Min problemformulering lyder,

Hur fungerar Heritrix[10] som arkiveringsmedium för webben?

Spänningen mellan tekniska och konstnärliga medier som strukturerar Kittlers *Nedskrivningssystem 1800 • 1900* (2012) ligger till grund för min gestaltning av Heritrix med e-plikten som kontrast. Min hypotes är inledningsvis att den principiella skillnaden mellan de två är att Heritrix är ett *tekniskt* medium som spelar in webben i sin helhet, inklusive vad som för människor kan te sig redundant eller som brus, precis som grammofonen spelar in röstens fysiologi inklusive diverse kroppsljud och bakgrundsbrus som inte är direkt signifikanta för människor. E-plikten är snarare en lag om att en specifik grupp aktörer ska producera och skicka in litteratur.

Jag operationaliserar min problemformulering till följande forskningsfrågor, med underfrågor:

1. Vad karaktäriserar Heritrix som **medium** (med e-plikten som bakgrund)?
   - Vad är dess epistemologiska konfiguration och specifika tidsperspektiv (cf. Ernst 2013, Parikka 2010)?
   - Hur fungerar Heritrix i nedskrivningssystem 2000?
2. Hur kan Heritrix **arkiv** förstås utifrån ett mediearkeologiskt perspektiv?
   - Vad är brus och vad är information? Vad är det ett arkiv över?
   - Vad skrivs ned, jämfört med e-plikten?
3. Kan en förståelse för inspelningsmedium som mer än neutrala databärare gagna **arkivvetenskapen**?

Varför?

Där finns flertalet anledningar till varför Heritrix som medium är ett högst angeläget studieobjekt:

1. kw³s insamling är den enda storskaliga insamlingen av den svenska webben som genomförts (undantaget Internet Archive, som har en viss täckning av densamma) och en förståelse för Heritrix är således angelägen av rent **historiska** skäl,

10 För att följa Kirschenbaum (2012:22) så ska jag nämna att jag under hela min insamling använt mig utav en Acer Aspire 5100 dator med 1024 MB RAM och en 1.4 Ghz processor, och på den har jag kört Debian GNU/Linux 7 ”Squeeze” som operativssystem. Jag använder mig av version 1.14.4 av Heritrix, eftersom detta är versionen de använder inom kw³ och eftersom den är mycket väldokumenterad – den senaste versionen av Heritrix när detta skrivs är annars version 3.2.0 som släpptes i januari 2014.

2. Heritrix samlar in webben i dess ”naturliga”, binära tillstånd i form av **HTTP-svar**[11], något som inte gjorts tidigare och inte kommer göras i och med e-plikten; frågan om medium är viktigt när e-plikten och kw³ ska utvärderas och jämföras
3. Heritrix arkiverar inte bara innehållet på webben, utan även webben som **infrastruktur** och medium i och med att den spelar in till exempel svarstider, ej hittade webbsidor etc. urskillningslöst, något som lär intressera framtida forskare och som utgör en kritisk forskningsfråga även idag. Enligt Kittler, som utgör en teoretisk mittpunkt i min studie, är innehåll som skickas via medieteknologiska kanaler i stort sett *funktioner* av mediets materialitet; ”vad vi än tänker står vi inte fria från de medier som möjliggör själva tänkandet” (2003:26),
4. Att studera **programvara och datorer som medium** har inte gjorts tidigare inom arkivvetenskapen – man har bortsett från programvarors och mediers relation till och makt över handlingarna, något som blir en än mer tydligt brist i vår datorcentrerade samtid

I mitt arbete med Heritrix kommer jag behöva arbeta ihop en verktygslåda med metoder och teorier, som i sig kommer utgöra ett bidrag till vetenskapen eftersom tidigare exempel är svåra att finna och kommer från andra fält. Detta behövs eftersom

1. inom arkivvetenskapen har medium/databärare mest behandlats ur ett (tekniskt) nyttoperspektiv, utan den teoretiska och metodologiska rikedom som finns att ösa ur **medieteorin**[12], vilket är ett problem eftersom ett av arkivsektorns viktigaste och svåraste uppdrag är att arkivera det medium vi kallar webben,
2. inom arkivvetenskapen har man heller inte använt sig av de perspektiv som finns att inhämta från **mediearkeologi** när det gäller webben som arkiv och programvara som medium för både inspelning och uppspelning, vilket likaledes är fattigt då exempelvis Heritrix som programvara lika lite som medium utgör en neutral aktör i sitt sammanhang, och eftersom webben består av globalt distribuerad programvara,
3. vad man dock har inom **arkivvetenskapen** är begrepp, erfarenheter och teorier som i kombination med ovan nämnda fält kan belysa Heritrix speciella karaktär som ”arkivbildare”[13] med allt vad det innebär för det skapade arkivet, och för insamlingsverksamheter såsom kw³.

I min uppsats har jag inte möjlighet att ge en holistisk, allomfattande bild av mitt studieobjekt, varför avgränsningar måste användas för att skulptera fram ett forskningsbart problem.

---

11 Ett HTTP-svar är det som våra webbläsare får när de frågar efter en webbsida, se https://en.wikipedia.org/w/index.php?title=Hypertext_Transfer_Protocol&oldid=595358158#Response_message [Hämtad 21 feb 2014]. Mer om detta nedan i resultat och analys-avsnittet.

12 Medie- och kommunikationsvetenskap – de delar som jag skurit ut ur detta spretiga fält för att använda i min uppsats presenteras nedan i metod- och teoriavsnitten.

13 Arkivbildare är egentligen den institution eller organisation som avlägger arkivet (i det här fallet borde kw³ eller KB vara arkivbildare för det arkiv som Heritrix lämnar efter sig) (cf. Geijer & Lenberg & Håkan 2013:17).

Avgränsningar

Jag har valt att skala bort följande områden från min studie, alla intressanta i sig:

- **subjektiva upplevelser**, jag vill studera Heritrix som *tekniskt medium* och kommer därför inte intressera mig varken för människorna som skrivit koden, eller de som använder den eller som medieras i dess arkiv – här faller alltså även olika typer av sociala perspektiv tyvärr bort, såsom politisk ekonomi och feministisk teori,
- **alternativa typer av insamling**, Heritrix kan användas för många olika typer av webbinsamling, men jag kommer endast genomföra och skriva om de filer som lämnas efter en fokuserad insamling av en specifik webbplats via HTTP med de flesta av spindelns inställningar lämnade på det förvalda,[14]
- **historik**, jag undviker traditionell narrativ mediehistoria[15] på grund av hur sådan pressar in mitt studieobjekt i en struktur (narrativen) som är det främmande, och istället skriva diagrammatisk (det vill säga beskrivande med ambitionen att likna kartor mer än text) om Heritrix som medium,[16]
- **juridik,** e-plikten kontra automatisk insamling via spindel innebär olika förutsättningar rent juridiskt både för insamlande institution, webbplatser som samlas in samt användare, men problematiken kring detta kommer inte att fördjupas mer än nödvändigt,
- **analys av källkod**, en intim förståelse och närläsning av den källkod som utgör programvaran vore intressant, dessvärre måste en sådan avgränsas bort då det vore för tidskrävande relativt till de insikter som kan nås. Jag fokuserar på de filer som en körning av spindeln efterlämnar eftersom de trots allt utgör en volym i värdinstitutionens arkiv, och dokumentationen som finns i överflöd räcker för att analysera programvaran,
- **djupare analys av e-plikten**, eftersom lagen är såpass ny, och eftersom den medieteknologiska infrastruktur som behövs inte är fullt på plats[17], så utgör e-plikten i min uppsats endast en bakgrund mot vilken jag gestaltar Heritrix.

Varje område som jag nämner ovan är värda forskning, och min förhoppning är att min studie kan bli kompletterad av andra studier som anlägger andra perspektiv. Jag avgränsar av praktiska skäl, för att göra min studie genomförbar, fokuserad och skarp.

14 Exempelvis skulle man kunna göra en bredare insamling av större delar av webben, eller en fokuserad på specifika ämnen eller konversationer på sociala medier (se Lomborg 2012). Heritrix har även möjlighet att samla in via andra protokoll, såsom *File Transfer Protocol* (FTP), vilket jag inte kommer utforska i min studie.

15 Dylika historiska uppsatser har skrivits, se exempelvis Ivarsson & Lundén (1998) samt Lasfargues, Martin & Medjkoune (2012).

16 Kritiken av mediehistoriska perspektiv går djupt i den medievetenskapliga nisch som kallas mediearkeologi (cf. Ernst 2013, Parikka 2007 och Gansing 2013 med flera), vilken utgör min främsta metodologiska och teoretiska inspiration – denna kritik kommer att relateras till min studie i teori- och metodavsnitten nedan,

17 Att man fortfarande är tidigt i utvecklingen av olika tekniska/praktiska lösningar för hur e-plikten ska fungera bekräftas i korrespondens [mottagen 1 maj 2014] med Boel Larsson, programansvarig för pliktfrågor på KB.

Data

Den data jag producerar för min studie kan delas in i *arkiv*[18], samt *dokumentation*. arkivet utgörs av processuellt (automatiskt) genererade digitala objekt i fallet Heritrix, och specifikationer när det gäller e-plikten, medan dokumentationen utgörs av textdokument (skrivna av människor).

Syfte och mål

Mitt huvudsakliga syfte med min uppsats är att den ska utgöra ett bidrag till arkivvetenskapen som utökar förståelsen för spindlar som medium. Utöver detta vill jag att min uppsats

- syntetiserar en metodologisk verktygslåda med element från medieteori, mediearkeologi som kan återanvändas, utvecklas och anpassas av andra, även inom arkivvetenskapliga undersökningar,
- enkelt kompletteras med forskning ur perspektiv jag valt bort, exempelvis antropologiska/etnografiska eller juridiska,
- kan användas av framtida forskare som vill förstå och studera de arkiv som efterlämnats av Heritrix (till exempel på KB) som arkiv över webbens materialitet och dagens nedskrivningssystem,
- kan användas av nationalbibliotek och andra som använder sig av Heritrix för att teoretisera och fördjupa sin insamlingsverksamhet,
- visar kw³s relevans efter lagen om elektronisk pliktleverans.

18 Ordet arkiv är problematiskt eftersom det betyder olika saker beroende på kontext. I min uppsats skriver jag *arkiv* när jag menar en arkivbildares samlade arkivlagda handlingar – alla de filer som efterlämnats från kw³ exempelvis – medan en specifik körnings av Heritrix efterlämnade filer benämns *volym*. Jag kallar .arc-filen som innehåller själva HTTP-svaren för *webbarkiv*, och jag undviker att använda ordet arkiv i Foucaults mening (jag skriver istället *nedskrivningssystem* och refererar då till Kittler). Slutligen undviker jag helt den datavetenskapliga förståelsen av arkiv som komprimerad fil.

Tidigare forskning

I detta avsnitt går jag igenom forskning som tidigare behandlat mitt studieobjekt. Jag använder mig av forskning från medie- och kommunikationsvetenskap, och istället för att presentera den här förlägger jag den till teorikapitlet. Mitt studieobjekt är strikt avgränsat till Heritrix snävt förstått som *medium*, och med den avgränsningen (som exkluderar exempelvis traditionell historia) kan man skönja två stora forskningsområden där spindlar figurerar,

1. Datavetenskap
2. Samhällsvetenskap

Tidigare forskning bortom arkivvetenskapen

Spindelalgoritmer, och implementationer av olika slag på en mer abstrakt nivå, utgör ett väl utforskat område inom datavetenskapen, inte minst till följd av att jätteföretag såsom Google tjänar sina pengar på sin söktjänst och därför är måna om att indexeringen av webben sker på absolut mest effektiva sätt. En artikel som går igenom olika forskningsprojekt som är direkt initierade av Google visar att alla projekten antingen handlar om spindelalgoritmer eller om olika typer av bearbetning av insamlade data (Cafarella, M., Chang, E., Fikes, A., Halevy, A., Hsieh, W., Lerner, A., Madhavan, J. & Mutukrishnan, S. 2008).

Inom digitalt långtidsbevarande (eng. Long-Term Digital Preservation, hädanefter förkrotat till *LDP*) och webbarkivering, som är underdiscipliner till datavetenskapen, behandlas spindlar pragmatiskt med frågor om hur arkiven ska bevaras och hållas tillgängliga på bästa sätt (Masanés 2006:177), samt hur man designar ett urval och sedan rent praktisk samlar in webben (ibid.:71-112).

Inom samhällsvetenskapen studeras användande av spindelalgoritmer som verktyg för att studera mänskliga diskurser på webben (Lomborg 2012), och mer generellt utgör spindelalgoritmer, Google och vad dessa innebär politiskt för samhälle och människa ett stort forskningsområde som jag inte går in på djupare här (jag kan nämna Terranova (2004)).

Tidigare forskning inom arkivvetenskapen

Inom arkivvetenskap har man inte behandlat Heritrix eller ens spindelprogram som arkiveringsmedium tidigare, och när det gäller medium i allmänhet finns där några områden som utforskats,

1. Praktisk forskning kring olika lagringsmediers hållbarhet och lämplighet, samt råd om migrering mellan lagringsmedier, speciellt vad gäller digital långtidsbevaring, exempelvis Dollar (1999), RA-FS 2009:1, Giaretta (2011) eller den forskningsöversikt som ges i Geijer et. al. (2013). Dylika frågor är inte intressanta i min studie eftersom jag behandlar Heritrix som medium och inte bryr mig om hur arkivet bör lagras,
2. Teoretisk forskning kring vad digitalisering av analogt material innebär, samt vad som utmärker digitala arkiv och hur det påverkar användare, exempelvis Latham (2010, 2011), Rojas (2009); Ernst (2013) behandlar digitala arkiv och historiografin. Forskningen beaktar mediers betydelse vilket är intressant, men konvertering från analogt sker ju inte med Heritrix så forskningen har inget större värde för min studie,
3. Forskning om hur nya digitala medier kan användas för att öka tillgängligheten till arkivmaterial och vad det innebär för användarens förståelse av kontext, proveniens etc., exempelvis Huvila (2008), Bak (2012) och Fear & Donaldson (2012). Det har på senare tid skett en svängning åt att fästa stor vikt vid användare av arkiv som medskapare av arkivet, speciellt med olika typer av webbapplikationer som tillåter större interaktivitet än tidigare mer stängda arkiv – Cook skriver att vi nu är på väg in i ett delvis nytt paradigm för arkivteoretiskt tänkande, där arkivarier blir aktivister som är integrerade i grupperingar och tillsammans med dem, med hjälp av digitala verktyg, skapar en ny typ av arkiv (Cook 2012:113). Kittler skulle väl snarare anse att det är nya mediasystem som skapar nya roller för ”arkivister”, och i sig är vad arkivarier tror sig göra egentligen inte intressant; jag gör en avgränsning här och följer Kittler,
4. Forsking kring immaterialrätt och arkiv, exempelvis Snickars (2011) och Iacovino & Todd (2007). För de som forskar i dessa ämnena lär min uppsats vara intressant, men jag går inte in på frågan själv.

På det stora hela finns där alltså inte mycket forskning inom arkivvetenskapen som är *direkt* relevant för min studie – dock så tror jag att min studie tillhör en framspirande del av disciplinen, som lär bli mer viktig framöver. Arkivarier kommer bara att få mer att göra med elektroniska medier – istället för att försöka ignorera dem eftersom de är komplexa, bör vi, anser jag, djupdyka och ta med oss vårt eget unika disciplinära kunnande till en teknisk materialitet där den annars riskerar att helt saknas.

Teori

Detta avsnitt diskuterar de olika texter och författare jag hämtar inspiration, verktyg och teori ifrån till min studie. Min litteratur går inte enkelt att dela upp i teoretisk respektive metodologisk, varför jag istället övervägande diskuterar teori och sedan sammanfattar om hur jag använder den i min studie. Först ger jag en *översikt* över min litteraturgenomgång, sedan följer epistemologiska och ontologiska grunder i form av Kittlers medieteori om *nedskrivningssystem.* Medieteorin blir sedan operativ som *mediearkeologi*, och jag avslutar med en *sammanfattning*.

Nedskrivningssystem

Nedskrivningssystem är den ontologiska kategori som är ankare för min studie. Jag börjar med att introducera *Kittlers medieteori* genom att beskriva hans läsning av Foucault och hur den skiljer sig från andra mer socialt eller kulturellt orienterade läsningar och hur den skisserar ett nedskrivningssystem. Jag kommer in på Claude *Shannon*s modell som Kittler hämtar mycket inspiration från och som är grunden för Heritrix operativa miljö webben, och genom att diskutera och svara på en del av den *kritik* som Kittlers verk utstått fördjupar jag teorin om nedskrivningssystemet och hur metoden för att teckna detta blir mediearkeologi, avsnittet som följer efter.

Kittlers medieteori

Michel Foucault med sin diskursanalys[19] förändrade de flesta socialt eller humanistiskt orienterade akademiska discipliner i grunden; olika inriktningar inom mediearkeologi och medievetenskaper i stort kan med fördel klassificeras utifrån hur de *läst* Foucault (Huhtamo & Parikka 2012:8).

Angloamerikanska medieteoretiker fokuserar på sociala och kulturella aspekter av medier där teknologin får sin mening och signifikans genom de diskurser den medverkar i (ibid.). Man läser Foucault som att diskurser är överordnade medier och bestämmer hur de senare kan användas och förstås. Diskurser, dvs. språket som ”ömsesidiga relationer, och det ständigt existerande avståndet mellan intentioner som relaterar till varandra” (Foucault 1984, citerad i Neumann 2003:25) är vad samhället består av, vad människor lever i och med (ibid.). Ett sätt att närma sig diskurser är att

19 I min uppsats berör jag inte närmare den uppsjö av olika typer och varianter av diskursanalys som florerar inom humanistiska ämnen, utan jag fokuserar på den variant av genealogi/arkeologi som kommer till uttryck inom mediearkeologin, specifikt den mer materialistiska versionen såsom den används av till exempel Kittler, Parikka och Ernst. Diskursanalyser som tar som sitt objekt de diskurser som faktiskt uttalas eller praktiseras (cf. Neumann 2003:17) är egentligen ointressanta till sitt innehåll, Kittler fokuserar istället på själva hårdvaran som gör diskurser och arkiv möjliga (2012:10); min första avgränsning (se *Avgränsningar* ovan) tar sikte utefter denna skiljelinje och skär bort diskursen.

tolka texter, ofta med hjälp av hermeneutiska metoder där man läser texter upprepade gånger för att ackumulera förförståelse som sedan används i nästkommande läsningar. På detta sätt vill man, iallafall traditionellt, komma åt den egentliga *meningen* bakom en text, vad författaren *ville säga*.

Hermeneutiker[20] gör för Kittler fatala misstag när de ser tolkningen och frammanande av människorna bakom diskurser som det enda sättet att närma sig någon verklighet – de har en tendens att ständigt tvivla på en yttre världs existens, deras ontologi består egentligen helt enkelt av subjekts kognitiva upplevelser av i grunden diskursiva fenomen (Olsen 2003:88). Som om hermeneutiken som metod eller medium självt inte har några materiella förutsättningar eller någon materiell historia!

Kittler vänder på det hela och skriver *mediets* genealogi, vilket kan te sig som teknikdeterminism men som är mer nyanserat än så. Istället för litteraturvetenskapens ständige protagonist Författaren (eller Människan) ser Kittler de olika apparaterna som gör varje diskurs möjlig och nedskrivbar som det primära, eller snarare *förbindelserna* mellan människan och medieteknologiernas materialitet (Fischer & Götselius 2003:12). ”Bakom de estetiska formerna och fantasmerna avtecknar sig ett »nedskrivningssystem»[*sic*], det vill säga ett »nätverk av teknologier och institutioner som tillåter en given kultur att utvälja, lagra och behandla relevanta data»” (ibid.).

Nedskrivningssystem är historiska och tenderar revolutioneras eller helt överspelas när ny medieteknologi införs, och två sådana skärningspunkter är vad som figurerar i Kittlers *Nedskrivningssystem 1800 • 1900* (Andersson, Fischer & Götselius 2012), ursprungligen skriven som habilitationsskrift i germanistik 1985. Boken blev nätt och jämnt accepterad som först efter hård debatt i betygskommittén, men utgör idag en klassiker bland mediehistoriker (ibid.:10). Boken handlar om medier och makt, och Kittler skriver 1800[21]-talet som ett ”informationssystem baserat på skriften och de nya praktiker som omger den, såsom allmän alfabetisering […] och hermeneutisk ordbehandling […] som syftar till att forma fulländade subjekt” (ibid.:11). I kontrast splittar 1900-talets experimentella psykovetenskaper den ”sensoriska varseblivningen i diskreta element och olika tekniskt definierade funktioner, vilket i sista hand upplöser subjektet” med hjälp av nya analoga medieteknologier (grammofonen, filmen och skrivmaskinen) (ibid.).

Kittler ser olika typer av mediesystem som delvis ihopkopplade i sitt samtida historiska nedskrivningssystem, och att de olika diskurser som produceras förutsätter och genererar varandra genom mediesystemen. Den litterära diskursen på 1900-talet fungerar endast som en inbäddad del av nedskrivningssystemet 1900, tillsammans med psykofysiken och psykoanalysen, och därför är definitioner av vad författare gör, vad som gör en text litterär etc. bestämda av nedskrivningssystemet som kommunikationssystem (Johnston 1997:4).

---

20 Hermeneutiker är knappast en sammanhållen skola eller någon entydig metod; när jag i min uppsats skriver om hermeneutik är det den Kittlerska eller mediearkeologiska tolkningen av vad som är hermeneutikens kärna som åsyftas; enligt den handlar det alltså om en tro på möjligheten att tränga in i en författares tankevärld genom att upprepade gånger läsa och tolka texter.

21 Genom att benämna nedskrivningssystemen 1800, 1900 och 2000 istället för till exempel romantik, modernism och postmodernism undviker Kittler kulturhistorisk självförståelse och hermeneutik som bara är i vägen när man vill se *media* (Fischer & Götselius 2003:15).

Diskurser är relationella nätverk av subjektspositioner från vilka en uppsättning praktiker och utsagor kan produceras, och dessa subjektspositioner sammanfaller inte med individer. När utsagor skrivs ned kan nedskrivningssystemet konturer och regler skönjas i själva praktiken/tekniken, och här håller Kittler med Marshall McLuhan kända devis om att *the medium is the message* (Johnston 1997:5). För Kittler handlar Freuds psykoanalytiska skriverier om nedskrivningssystemet 1900, och det är därför Freud måste föregripa anklagelser om plagiat när det visar sig att den schizofrene domaren Schrebers nedtecknade vanföreställningar på det stora taget är samma analys som den som Freud själv fört fram (långt senare) i sin libidoteori; ”[k]larare fall av litterär vittnesbörd [om att de båda utgör funktioner av samma nedskrivningssystem från olika håll] finns inte” (Kittler 2012:414).

För Kittler som för Foucault är diskurser praktiker som ”systematiskt formar de objekt som [framställs]” (Fischer & Götselius 2003:21) och diskursanalys handlar därför om att synliggöra de ”makt-, minnes-, överförings-, och bearbetningsapparaturer som etablerar villkor för vad som kan sägas i en viss kultur vid en viss tidpunkt” (ibid.). Fokus ligger på kanalerna genom vilka signaler överförs, och nedskrivningssystemet är överordnat diskurserna som produceras.

Shannon

Claude Shannon lade grunden redan på 1940-talet för hur dagens informationsteknologiska infrastruktur i allt väsentligt skulle fungera med sin matematiska teori om kommunikation, som senare applicerades och blev kommunikation mellan datorer (Lindgren 2009:11). Redan 193 visade Shannon att hela den boolska algebran[22], och med den möjligheten att lösa varje numeriskt logiska problem, går att implementera i vanliga telefonväxelsystem, sedan var det bara fråga om att bygga chip för att förminska sådana för att sätta igång datorns utveckling mot medium på allvar (Johnston 1997:153). För att förstå Heritrix och dess operativa miljö (webben) är det därför vara av intresse att lägga fram Shannons diagram,

22 Boolsk algebra är algebra där variablar endast antar värdet falskt eller sant (1 eller 0), och där matematiska operationer går att reducera till *OCH*, *ELLER*, samt *INTE* – boolsk algebra är fundamental för digital elektronik, och en dator arbetar i grund och botten med en lång räcka ettor och nollor med hjälp av enkla operationer – bara väldigt snabbt. Se https://en.wikipedia.org/wiki/Boolean_algebra

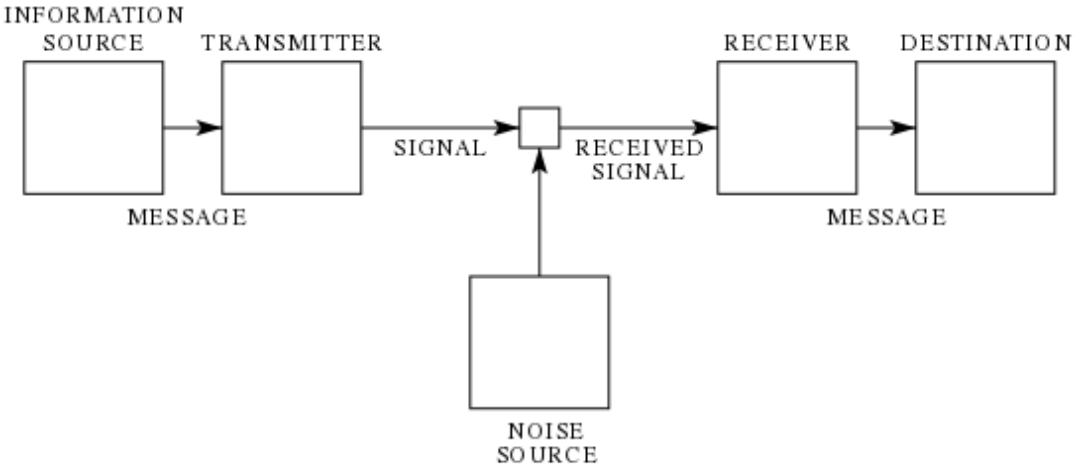

Fig. 1—Schematic diagram of a general communication system.

(Shannon 1948:2).

Vad som är grundläggande för Shannon är att han använder begreppet information på ett speciellt sätt, som inte har något att göra med *mening* – eller snarare, eventuell mening som skickas genom ett kommunikationssystem påverkar inte *informationen* som skickas. Självklart gäller dock det omvända, kommunikationssystemet kan väsentligen påverka eventuell mening i information (Weaver 1998:8), exempelvis så kan ord missförstås när de talas över en skakig telefonlinje. Egentligen gav Shannon aldrig en definition av begreppet information, utan endast en modell för hur man ”kvantitativt kan mäta reduktionen i osäkerheten att mottaga en kommunikation” (Byfield 2006:127), ett mått som han sedan kallade information (ibid.).

Denna modell gäller i grunden även för de protokoll och system som utgör webben och Internet – även om de senare är extremt komplicerade av många års förfining och applicering. Jag återkommer till Heritrix operativa miljö i min resultatanalys nedan.

Kittler använder Shannons matematiska modell för kommunikationssystem, där medier primärt är kanaler för signaler, och där meningen precis som hos Shannon inte utgör en intressant del av begreppet information (Weaver 1998:8). Kritiker menar att Kittlers teorier endast är applicerad Shannon (Hansen, citerad i Hayles 2005:34), mycket på grund av hans futuristiska scenario där våra digitala mediesystem innebär en överspelning av människans roll i kunskapscirkulation – istället sammankopplas mediesystemen och ”det absoluta vetandet gå[r] som en oändlig slinga” (Fischer & Götselius 2003:34). Så fortsätter vi in på kritiken.

Kritik

Vissa futuristiska överdrifter står ut när man läser Kittler, och för sin ”idiosynkratiska kosmologi” (Kirschenbaum 2012:6) har han blivit hårt kritiserad även av dem som annars tar till sig vissa av hans grundläggande intressen och insikter. Enligt Kirschenbaum lägger Kittler fram en teleologisk narrativ om fullständig mediakonvergens och slutet på historien (ibid:7), en berättelse som påminner starkt om Francis Fukuyamas neo-konservativa och ökända essä ”The End of History” (Gansing 2013:70).

Vissa angloamerikanska teoretiker avfärdar Kittlers ”mediateori utan människor” (Gansing 2013:37) som verklighetsfrånvänd och föredrar istället sina egna prioriteringar, där exempelvis människors vardagsliv kring medier utgör fokus (ibid.). Kittlers sammankoppling av olika mediesystem figurerar även här, eftersom tendensen är svår att ignorera – man talar inom mer socialt och kulturvetenskapligt orienterade riktningar om ”media convergence” (ibid.:46), och även inom kritiska riktningar uppmärksammas den vertikala integrationen av olika media och den kapitalistiska organiseringen av vår sensoriska omvärld (ibid.:47).

I sin bok *My Mother was a Computer: Digital subjects and literary texts* jämför Hayles (2005) Kittlers syn på ny media med Mark B. N. Hansens, som får representera en motpol. Hansen väljer att se ny mediateknologi, som en reaktion på vad han upplever som Kittlers teknofetishism, exklusivt genom dess effekter och dess samspel i kroppsliga subjekt i form av användare och tittare, och som i stort sett oberoende underliggande medium (Hayles 2005:35).

Hayles själv går en mellanväg mellan dessa två extremer, och anser att de båda teoretikerna ”sitter på varsin sida av en gungbräda och försöker trycka på sin egen sida, utan att inse att deras ändar är sammankopplade i mitten” (ibid.). Självklart, skriver hon, beror mediateknologier på dess användare och tittare – om inte annat så designar man ju medvetet medieteknologi för mänsklig tillgänglighet, och utan människor hade ingen kunnat uppfinna medier; å andra sidan determinerar medieteknologi inte bara receptionen av innehåll utan de formar även våra kroppars förmågor och sinnen (ibid.:36). För mig blir valet av placering på gungbrädan ett metodologiskt val mellan att undersöka medieteknologi och att undersöka dess mänskliga sändare, mottagare och uppfinnare etc. - jag väljer att placera mig hos Kittler och avgränsa bort de som sitter åt Hansens håll, även om jag inser att de finns där, och att de gärna får hjälpa till i undersökningen av Heritrix som medium i framtida studier.

Om man förbiser stilistiska särdrag framgår det att Kittler inte hävdar att datorer skulle innebära slutet för litteraturen eller den västerländska civilisationen utan endast att litteraturens roll som mediehistorisk huvudroll är över; det betyder förstås inte att litteraturen kommer försvinna, men den har blivit ett medium bland andra och formar inte vår samtids villkor på samma sätt som den gjorde tidigare (Fischer & Götselius 2003:31). Ernst går längre i sitt förkastande av historiska eller narrativa metoder till förmån för datorteknologins tidsperspektiv och faller kanske mer tydligt i en ”slutet på historien”-fälla (Gansing 2013:70) – men även här vore det en onyanserad överdrift. Litteratur är för Kittler uppenbart fortfarande i många fall ett potent analytiskt instrument – annars hade han väl inte skrivit böcker. (Och annars hade han väl inte så ofta refererat till böcker i sina böcker.)

För Kittler innebär diskursanalys arbete med *minnesersättningar*, medieteknologi för att motverka glömska, *i sina specifika former* och utan att falla tillbaka på en metafysisk urskrift bortom dem, ett taktiskt misstag som till exempel den franske filosofen Jacques Derrida enligt Kittler gör (2003:58-59). Den filosofiska eller metafysiska diskursen ”vill inte veta av att minnesfunktioner kan stängas av och förstöras” (ibid.:59) och den har svårt att se sig själv som en ickeautonom diskurs/institution bland andra. Istället för att producera utsagor om diskurser som

”trancendentalt och kategorialt” (ibid.) existerar som friflytande text, tolkar Kittler Foucaults diskursbegrepp som något som har en materiell basis i mediesystem, och som begränsas eller till och med utgör en funktion av ett nätverk av mediaapparater– vanligen översätts nedskrivningssystem till just *discourse network*, på tyska heter det dock *Aufschreibesysteme*, en term som liknar den svenska översättningen mer än den engelska.

Den tekniska definitionen av minne är ”en apparat i vilken man kan introducera information och sedan extrahera den vid en viss senare tidpunkt” (Ernst 2013:100) och det är denna typen av definition, istället för de mer socialt orienterade, som är grundläggande för Kittlers och mediearkeologers studie av media; denna definition beskriver bättre de apparater som nedskrivningssystem utgörs av – de andra typerna av minne, såsom kulturminnen eller personliga minnen, kan ses som *inkopplade* till ett nedskrivningssystem. Media blir på så vis den primära (men inte enda) historiska drivkraften.

Denna position, att media utgör en ”historisk aktör” (Robertsson 2011:2), är i sig inte superkontroversiell och figurerar i vad mediearkeologer kanske skulle förkasta som angloamerikansk mainstream mediahistorik, till exempel i antologin *Media History and the Archive* redigerad av Robertsson (2011). Till exempel skriver nyhetstidningshistorikern John Nerone att vad redaktörer och journalister tror att deras tidning handlar om inte sammanfaller med vad ett nyhetsmedium *faktiskt* handlar om (2011:21) och att nyhetsmediet kan ses som ett komplext nätverk eller system (ibid.:23) vilket ju åkallar Kittler även om denne inte refereras. Paddy Scannell slår i sitt bidrag fast att ”all teknologi har pragmatiska funktioner som formar och strukturerar spelrummet och skalorna i vilket den används” (Scannell 2011:44). I sitt slutord till antologin försvarar John Durham Peters Kittler – han är måhända strukturalist och ”svuren fiende till det liberala subjektet” (2011:113) men det gör honom inte till teknikdeterminist, utan kanske i värsta fall till cyniker.

Kittlers diskursanalys ”praktiserar en 'glad positivism' [och] handlar inte om den gemensamma urskrift som all närvarometafysik förutsätts ha glömt, utan om flera gömda tekniker som utformats för att motverka glömska. Den avstår från kommentarens njutningar […] för att gripa sig an de arkiverade texterna med en arkivaries tålamod och brist på läslust” (Kittler 2003:59). Just sådant tålamod och sådan brist på läslust (i meningen att önska läsa innehåll) är vad jag försöker praktisera i min studie – jag åsidosätter arkivets egentliga innehåll för att fokusera på formen. Samma tålamod och brist på läslust karaktäriserar förövrigt även Heritrix (och datorer) som medium – det som vi ser som (möjligen upprörande eller intressant) innehåll online som visas på skärmen är precis likadana ettor och nollor som allt annat för en dator, och kan även manipuleras enligt samma booleska algebra.

Diskursbegreppet, det nedskrivna språket, är alltså inte en ”räcka tecken i tal eller text, vilka inom sig bär en mening som det är dess syfte att kommunicera” utan ”en serie händelser som förutsätter en viss materialitet, en viss regulativ karaktär, en viss diskontinuitet etcetera […] en praktik som systematiskt formar de objekt som den framställer” (Fischer & Götselius 2003:21). Därför signalerar de tekniska mediernas[23]

23 Kittler ser tekniska medier som medier som registrerar en specifik del av sin omvärld utan omväg via en människas hand – fotografi, exempelvis, är ett tekniskt medium som låter ”naturen”

intåg på 1900-talet ett avgörande nederlag för alla försök till hermeneutisk förståelse som det primära sättet att tillgodogöra sig skriven text, ett förhållande som enligt Kittler utmärkte nedskrivningssystemet 1800. Nedskrivningssystem 1900 överger all mytologi om att hermeneutiskt ”tränga in i författarnas ande [genom texter]” (Kittler 2012:386) och reducerar istället text till olika typer av psykofysiska och psykoanalytiska modeller. Kittlers diskursanalys är präglad av detta skifte, och för att sedan förstå 1800-talets nedskrivningssystem var en analys av psykofysiken/analysen med hjälp av metoder och apparater från 1900-talet nödvändigt. Vårt samtida nedskrivningssystem kräver också andra metoder än hermeneutiska för att träda fram, och mediearkeologin framstår här som ett försök att metodologiskt framskrida emot nästa skifte för att kunna se tillbaka på nedskrivningssystem 2000.

I min studie blir nedskrivningssystem ett centralt begrepp på grund av hur Kittler tecknat nedskrivningssystemet 2000[24] med den digitala datorn som centralt och allomfattande medium, som likt skriftmonopolet i 1800-talets nedskrivningssystem i slutändan gör konceptet medium osynligt och meningslöst,. Alla medier kan simuleras och reduceras till samma typ av inskription (binär kod i en dators minne) och i den logiska slutändan kan dessa kopplas ihop till en enda rundgång där människan inte längre ingår i produktionen, cirkulationen eller ens receptionen av kunskap (Kittler 2010:11-12) – lyckligtvis ”»härskar partiellt sammankopplade mediesystem»” (Kittler 2003:34) fortfarande, och ett prominent och viktigt medium som är inkopplat är webben, i sin tur inspelat av Heritrix.

Heritrix är ett medium för datorer, eller för de nätverk genom vilka datorer kopplas samman med varandra, vilket innebär att framtida studier av nedskrivningssystem 2000 kan i dess arkiv inte bara studera representationer av webben utan till och med en nära på *optisk* inspelning av den. För den som vill förstå nedskrivningssystem 2000 är även Heritrix själv som medium av intresse.

För att sammanfatta, Kittler används i min studie

1. för att motivera och specificera mina val inom avgränsningar och metod,
2. för att specificera vad jag letar efter i Heritrix (nedskrivningssystemet 2000)

Kittlers perspektiv på medier fattas inom arkivvetenskapen, och alltså har vi redan nu kommit en bit på vägen emot en mer eftertänksam och skarp arkivteoretisk syn på (speciellt nya) medier. För att göra teorin mer användbar krävs dock ytterligare utveckling, och en teoribildning som växt fram åtminstone delvis ur Kittlers arv är mediearkeologin, ämne för nästa del av teoriavsnittet.

---

avteckna sig via en kemisk reaktion på en plåt, medan människors subjektiva val är involverade på ett avgörande sätt när man använder mediet penna och papper för att rita av något.

24 Det följer av Kittlers behandling av tidigare nedskrivningssystem att man inte kan på ett tillfredsställande sätt beskriva det nedskrivningssystem man utgör en funktion i – innan tekniska medier introducerades kring 1900 kunde man inte första hur litteraturen fungerade som monopolkanal i nedskrivningssystemet 1800 (cf. Kittler 2012:329). Det omöjliga men ändå nödvändiga man kan göra är därför ”tentativa försök att närma sig en komplex men livsviktig problematik som först när den är historiskt överspelad kommer att framträda med tydligare konturer” (Fischer & Götselius 2003:20).

Mediearkeologi

> Archaeologists should unite in a defense of things, a defense of those subaltern members of the collective that have been silenced and ”othered” by … imperialist social and humanist discourse … This story is not narrated … , but comes to us as silent, tangible, visible and brute material remains (Bjørnar Olsen, citerad i Sobchack 2011:323)

Detta citat ger an klangen till vad som är en del utav mediearkeologins grundläggande *ethos*, materialistisk, antinarrativ antihermeneutik (Sobchack 2011:323), som tar som sin utgångspunkt medieteknologier i dess allra vidaste bemärkelse, inklusive imaginära och/eller omöjliga medier (cf. Kluitenberg 2011). Kombinationen av medieteori och en arkeologisk ambition som en utveckling och operationalisering av den teori som jag mejslade fram med hjälp av Kittler ovan.

Men mediearkeologin är ett komplicerat verktyg – vad mediearkeologi handlar om har snart sagt varit upp till mediearkeologen – disciplinen är odisciplinerad, nomadisk i att den rör sig mellan och transversalt[25] genom olika andra vetenskapliga fält, och med en vid uppsättning metoder och teorier till sitt förfogande. Nu ska jag berika arkivvetenskapen med den, men först en översiktlig historisk tillbakablick som är tagen till stora delar ur Gansings (2013) teoriavsnitt samt ur antologin *Media Archaeology: approaches, applications and implications* redigerad av Huhtamo och Parikka (2011). Dessa två källor balanserar varandra bra – Gansing har i sin doktorsavhandling ställt sig kritisk till olika delar av mediearkeologin, även om han senare anammar sin egen variant, Huhtamo och Parikkas antologi ger en uppsjö olika exempel på hur disciplinen praktiseras samt även analys av disciplinen självt.

Sedan år 2000 har medieteoretiska arbeten med en historisk orientering blivit populära att skriva (Gansing 2013:61). Det var ur denna vändning som termen mediearkeologi började användas, först av Siegried Zielinski, som anser att metoden genom att ”öppna upp heterogena mediehistorier motarbetar standardisering” (Zielinski, citerad i Gansing 2013:61) och att den erbjuder ”verktyg för att gräva ut hemliga gångar i historien, som kan hjälpa oss att finna vår väg till framtiden” (ibid.); den första artikeln som han skrev i ämnet handlar om artistsubjektets relation till nya medier under slutet av 1900-talet (Gansing 2013:61). Disciplinen har sedan spridit sig och antagit olika former.

Mediearkeologi som posthumanistisk teoribildning har som en av sina grundpelare en kritik av den litterära och historiografiska narrativa framställningen, en kritik som inspirerats av Hayden Whites *Metahistory* (Parikka, i Ernst 2013:30). Zielinski arbetar i sitt mest kända verk (2006) genom att öppna snitt ner i mediehistorisk djuptid, ett koncept som refererar till den hisnande tidsrymd som geologin upptäckte när bevis framkom att jorden var långt äldre än 2000 år. Genom att gräva fram bortglömda uppfinnare, trollkarlar och andra med deras olika idéer (emellanåt realiserade) om möjliga och omöjliga medieteknologier vill han skriva radikalt

25 Transversalitet är centralt för Gansing och syftar generellt på ideér som löper tvärs över/igenom specifika situationer och teorier (2013:15). Transversalitet är ett derivat av adjektivet transversal som härstammar inom geometrin och betecknar en linje som skär genom ett system av linjer – inom kulturvetenskap står konceptet för en rörelse som går ”över och bortom territorium eller institutioner och deras givna praktiker (från en plats och ett syfte till ett annat), som en utmaning av för givet tagna strukturer och system genom att länka samman heterogena element” (ibid.:17).

heterogena historier och samband (Zielinski 2006). Ambitionen är att utöva motstånd mot normaliserad och ”mainstream” medievetenskap (Huhtamo & Parikka 2011:10) – på senare tid bedriver Zielinski vad han kallar *variantologi* som ett sätt att än en gång undgå att hans metodik assimileras in i vanlig medievetenskap eller in i akademin som sådan (ibid:12).

Variantologi genomför ”'lokala'” (ibid.) undersökningar och vägrar acceptera eller använda sig av övergripande historik och generella förklaringar. Zielinskis motstånd mot systematisering och teoretisering samt hans fokus på mediearkeologisk djuptid ter sig egenartad och inte speciellt relevant för min studie, om man nu inte vill kalla kopplingen mellan etologins teorier om svärmen, insekter och programvarors beteende (ett tema som utforskas av Parikka (2010)) för ett snitt ner i djuptid.

Motstånd mot etablissemanget ger mediearkeologi en epistemologisk koppling till emancipatoriska, kritiska teoriskolor där narrativa framgångssagor, där teknologisk utveckling går linjärt från bättre till bättre, ses som högst suspekta (Alvesson & Sköldberg 2008:304-307). Detta kan om inte annat kan stå som ett motargument mot kritik om att speciellt Ernsts teori går att reducera till apolitisk hårdvarufetischism (cf. Parikka 2013). Konsumtionskapitalismen utgör i allra högsta grad ett studieobjekt även för mediearkeologin (Hertz & Parikka 2012:427), även politik inte utgör fokus i min studie så är det inget som ignoreras eller tas för givet inom mediearkeologin som sådan.

På liknande sätt som Zielinski analyserar Machiko Kusahara en zoetrop[26] från 1910-talets Japan som kallades ”Baby Talkie” och som när den placerades ovanpå en grammofonspelare var tänkt att fungera som den tidens hemmabio – rörliga bilder i fullfärg till ljudet från skivan som spelas på grammofonen (Kusahara 2011:123). Kusahara framkallar i sitt studieobjekts samtida Japan och den nationalism som sedan drog över landet efter den korta period av öppenhet mellan 1910 och 1930 som kallas Meijirestorationen; det förflutna blir samtida när en artefakt används på detta viset, en ”kortslutning” mellan dåtid och nutid som Ernst säger (2013:57).

Ernst har en rent antihistorisk eller antinarrativ ambition, och hans typ av mediearkeologi går längre än andra mediearkeologer i sin kritik av att använda en historiografisk modell för sitt skrivande, alltså en där man ordnar händelser kronologiskt och enligt en narrativ (2013:56). Genom att undvika den makrotemporala världen och istället fokusera på maskiners mikrotemporalitet går Ernst bortom både Kittler och Foucault, och idealet är att registrera omvärlden likt en kamera, med maskinens kyliga, ej diskursivt orienterade blick som mäter och beräknar snarare än tolkar (ibid.:59). I min uppsats har jag till viss del anammat Ernst kritik som stilspråk exempelvis i form av att försöka ordna saker i listor istället för i kronologi – men jag tror inte att stilspråk är det viktiga att ta med sig från Ernst, utan en förståelse för vad som utmärker tekniska medier, där hans ontologi är användbar.

26 En zoetrop, uppfunnen första gången så tidigt som 180 e.Kr. i Kina, ser ut som en liten burk vars insida är täckt med en serie bilder, och vars sidor är perforerade med små springor. När man snurrar zoetropen och tittar in genom springorna uppstår illusionen att bilderna rör sig. Se https://en.wikipedia.org/wiki/Zoetrope [Hämtad 8 mars 2014].

Gansing frågar sig dock var man egentligen kan hitta denna teleologiska, narrativa, optimistiska och evolutionärt skrivna mediehistoria, som mediearkeologi ska vara ett botemedel mot, egentligen? (2013:70). Om man placerar mediearkeologin som ett av många perspektiv på teknologisk utveckling, vilket är vad Gansing gör i sin bok, så framstår bilden som mer komplicerad och med mediearkeologi som inte speciellt inkompatibel med andra perspektiv (ibid.:71-76).

Vad Gansing tar med sig från mediearkeologin är istället främst konceptet om *transversalitet* som ursprungligen hämtats från Foucault (ibid.:79). En mediearkeologisk syn på teknologisk utveckling är inte främst ej linjär utan ej evolutionär, den skär tvärs igenom ett evolutionärt perspektiv för att generera nya insikter, skär ner till *produktionsplatsen* för diskurser/makt istället för att fokusera på de producerade institutionerna (ibid.). I den meningen vill jag se min studie som transversal, jag skär ner till *protokollen*, de epistemologiska konfigurationerna, som bestämmer hur mediet Heritrix funkar.

Som sådan kan mediearkeologens ”historiska fantasi” enligt Hayden Whites typologi (Sobchack 2011:328) sorteras som tillhörande en *romantisk* typ av historik, där gamla obsoleta medier återuppstår eller visar på nyheter i dagens situation, där det förflutna grävs fram ur dammet och återfår glans och vikt. Disciplinens andra historiografiska djupstrukturella tendens är att ”identifiera unika karakteristika hos objekt som befinner sig i det historiska fältet” (ibid.:329), alltså ett fokus på *form*. Som vi sett så är dessa karaktäristika bara ytligt beskrivande; viss mediearkeologi vill inte vara historisk alls (Ernst), och den mesta är iallafall djupt skeptisk mot historiska narrativa framställningar. Det är ett fokus på hur ting existerar bortom människors konceptuella värld – materialism – som är den primära tendensen.

Parikka ser sin mediearkeologiska epistemologi som en blandning mellan Gilles Deleuze och Kittler – de källor som Parikka använder i sitt vetenskapliga arbete konceptualiserar han (inspirerad av Deleuze) som *flöden* kopplade till abstrakta maskiner eller människa-maskin-assemblager (Deleuze kallar dem *maskin-phyla*) som cirkulerar eller produceras i Kittlers mer konkreta nedskrivningssystem. Genom att kombinera Deleuze rörelse- och emergensfokuserade filosofi med Kittlers rigida strukturer vill Parikka skapa ett spektrum inom vilket man kan röra sig med källor och material för att utlösa händelser och skapa nya insikter (2007:18).

Jag fokuserar på nedskrivningssystemet och mediet självt, så Kittler är egentligen nog epistemologisk bas för mig, men vad jag kan ta från Parikkas metodologi är dess fokus på att inte analysera studieobjektet som passiv källa, utan att *använda* eller aktivera det; jag inte bara genomför en insamling med Heritrix, utan använder resulterande volymer för att närma mig en del av nedskrivningssystem 2000. Likt Ernst vill jag plocka isär Heritrix och dess volymer och lägga fram delarna (2013:12).

I min studie fungerar mediearkeologin i att den

1. skärper fokuset på *form* genom att lokalisera den i tekniska mediers kyliga existens
2. är aggressivt inriktad på att generera ny kunskap genom att *använda* tekniska medier

3. skär transversalt via tekniska protokoll till mediet

Hur kan man då passa in den mediearkeologi jag konstruerat här i ett arkivvetenskapligt perspektiv?

Arkivvetenskapen, Kittler och reflektioner kring att skriva uppsats

Den viktigaste datan i min studie utgörs av digitala objekt som efterlämnat processer iscensatta genom att köra programvara. Med samhällets snabba utveckling mot e-tjänster, digitala arkiv och webbplatser för alla, påverkas även arkivvetenskapen i grunden. Klart är att ”arkivvetenskapen måhända tillhör den grupp vetenskaper som i alldeles särskild utsträckning påverkas av informationsteknologin (IT)” (Burnell 1995:100).

Burnell skrev tidigt om denna överlappning och vad den kunde tänkas betyda för arkivverksamheter (1995), och systemvetenskapens betydelse för datorteknologi och för texter behandlar även Hayles (2005) och Parikka (2007, 2010). Mer praktiskt orienterade är de systemvetare som ingår i min litteratur (Quisbert 2008; Giaretta 2011; Lindberg 2009) - Kirschenbaum behandlar lagringsmediets materialitet (2012). Den viktigaste skillnaden mellan de artiklar som exempelvis Dagens Nyheter skickar in till KB i standardiserat format för att uppfylla e-plikten och samma artikel från http://www.dn.se/ som den blir insamlad av Heritrix kan illustreras av en ersättning av litteraturvetenskap med mediearkeologi.

Det är här som mitt teoretiska perspektiv fyller en lucka i arkivvetenskapen. Mediets betydelse och makt är tydlig för Burnell, Giaretta med flera, men den mer praktiskt orienterade arkivvetenskapliga eller datavetenskapliga analysen kan eller vill inte se *varför* eller *hur*. Arkivvetenskapen har ur mitt perspektiv en något naiv världsbild. Som slutet på litteraturvetenskapens romantik har vi sett att tanken om ett verk, en författares tanke som kan skönjas eller hallucineras mellan raderna i en text, kan behandlas som just ett hjärnspöke – nedskrivningssystemet 1900 såg till att tidigare epoks högt vördade filosof-författare och läsare blev till *tjänstemän*, som ”jagar fram genom världen och ropar – sina meningslösa budskap till varandra” (Kafka, citerad i Kittler 2012:478). Informationen, i den kommunikationstekniska meningen[27], som sänds i kanalerna är i princip en effekt av kanalerna, litteraturen förblir ”just det avfall som den själv beskriver” (Kittler 2012:477). Arkivvetenskapen har inte uppfattat detta.

Kittlers magstarka formuleringar omformar jag till att man för att studera nedskrivningssystem måste frångå humanistiskt tolkande; att börja tolka exempelvis Heritrix manualer eller volymer som texter för att utröna sanningar om dess författare och deras diskursiva sammanhang, är lönlöst när man är ute efter att avtäcka de mediekonfigurationer och relationer som ligger bakom dessa uttryck. Och i mitt teoriavsnitt har jag velat visa att detta är en viktig målsättning att ha.

27 Claude Shannon uteslöt som vi sett ur sitt informationsbegrepp varje mening eller budskap som en irrelevant sidoeffekt, för att kunna kvantifiera och behandla information statistiskt; vad som spelar roll är huruvida en räcka bitar som stoppas in i en ända av en kommunikationskanal kommer ut i andra ändan i samma ordning (1963).

Mot detta kan man invända att det ter sig naivt att tro att jag kan närma mig ett studieobjekt utan att tolka, och det är det ju förstås – någon form av tolkning är omöjlig att undkomma när man skriver text. Annat vore det om jag kunde använda ett tekniskt medium och till exempel presentera min insamling av KBs http-svar som min studie, det hade varit en form av resultat utan egentlig tolkning. Nej, det handlar snarare om att ha ambitionen att anlägga ett annorlunda perspektiv än hermeneutisk eller historisk tolkning. Mediearkeologin kan ses som inspirerat av maskiners temporala värld, i vilken exempelvis rekursion, cyklisk och manipulerad tid råder (Ernst 2013:19). Det handlar om att tänka om sina förutfattade meningar om vad som är intressant hos ett studieobjekt; att tänka media utifrån de ”mönster, pulser och intervaller” (ibid.:18) som information existerar som innan de når någon form av manifestation som kan nå mänskliga sinnen.

Precis som Hayden White uppmanar historiker att skriva historia med en viss *självironi*, eller kritisk medvetenhet, om de narrativa strukturer och troper de formar sina källor enligt, anser Ernst att mediearkeologen också måste visa självironi när hen ”koncentrerar på de ej diskursiva elementen i det förflutna: inte på talare utan på maskiners makt/inverkan” (2013:45, min översättning) – man ska inte vara naiv åt endera hållet.

Det har för arkivvetare liksom för alla mänskliga användare blivit nödvändigt att förlita sig på algoritmiska representationer och transformationer för att använda digitala arkiv (Bazerman 2012:386), den enkla sanningen är ”att man inte kan se bitar[28] […] de är fullständigt onåbara för människans sinnen” (Levy, citerad i Kirschenbaum 2012:30, med min fotnot och min översättning). Att dessa representationer skall vara möjliga, verifierbara och pålitliga är det problem som står framför arkivvetare och systemutvecklare, och därför framskrider nu olika standardiseringsprojekt såsom OAIS (Giaretta 2011:47) och InterPARES (Giejer et. al. 2013:106-108). En mer avancerad teoretisk förståelse för mediet vore till gagn!

En mediearkeologisk uppluckring av tidigare självklara perspektiv kan visa hur medier strukturerar vår perceptuella värld på dessa för oss onåbara platser, mikrochippens värld. Tänk dig exempelvis hur strömmande videoklipp sända över en långsam uppkoppling bestäms av datorers och internets mediala villkor, vi ser mediets konturer bakom innehållet i form av *lagg*[29], långt innan semantikens inträde stammar strömmen fram sitt medium (ibid.).

McLuhan ansåg att innehållet i ett medium i princip alltid är ett annat medium (ett tv-program visar bilder som avbildar något, och till detta kanske vi hör en berättarröst som i sin tur återger en narrativ etc.); en mer utvecklad form av detta resonemang använder Hayles (2005), som har tagit Kittlers fokus på hur medier bestämmer diskurser till sig men komplicerar det hela genom att tänka sig subjekt och text som *ihoptvinnade* eller *sammanblandade* och att olika medier inte bara innehåller varandra utan står i ett *intermedialt* samband med varandra på ett komplext sätt (ibid.:7).

28 Ordet *bit* är en förkortning av *binary digit*, en binär siffra, 1 eller 0. När jag skriver bit eller bitar menar jag bit i denna bemärkelse, och inte bit som i ”bitar av en kaka” eller ”bit mig i fingret”.

29 Från eng. *lag*, i betydelsen fördröjning eller stakning

Heritrix webbarkiv innehåller förstås mediet webben, men relationen är mer komplicerad än så. Olika gränssnitt kan representera webbarkiv, exempelvis Internet Archives *Wayback Machine*[30] vilken tillåter att man surfar i webbarkiv på ett liknande sätt som vi annars surfar på internet, medan ett enkelt textredigeringsprogram kan återge webbarkiv som långa textfiler där webbplatser är arrangerade i den ordning de samlats in. Andra programvaror kan generera olika typer av statistik eller visa sidor som uppfyller vissa kriterier, vissa typer av filer etc. I min uppsats har jag avgränsat bort subjekt/innehåll och därför kan jag inte dra speciellt mycket nytta av Hayles intermedialitet, men den utgör ändå en viktig baktanke som förhindrar att man drar för snabba slutsatser om olika mediers relation till varandra. Shannons abstrakta mätinstrument för information är nödvändigtvis kopplad till *media* – vad människor upplever är bortom Shannons informationsbegrepp och skapas av media (Byfield 2006:128).

Mitt teoretiska perspektiv vill alltså fortsätta där andra stannar. Riksarkivet vill av praktiska och juridiska skäl behandla arkivhandlingar som om medium inte existerade eller var neutrala databärare; de delar av arkivvetenskapen som hör hemma här är alltså inte relevanta. Juridisk forskning sänds även den genom nedskrivningssystemet, vilket exempelvis Vismann skriver om när hon lägger fram akten som medium för administrationer, och hur det styrt dem och deras institutioner (2008). Därför utgår jag från *arkivet,* från den faktiska medietekniska apparat som föreligger, och för att operationalisera Kittler lade jag till en mediearkeologisk lins.

Rent dispositionsmässigt tar min teori uttryck i att min analys är organiserad kring epistemologiska konfigurationer för att beskriva grundläggande principer/teknologier för hur information överförs och bearbetas i tekniska medium, samt hur kunskapscirkulation är möjlig med mediet. Mitt teoriavsnitt är nu direkt kopplat till och svar på min tredje forskningsfråga, där jag efterfrågar en mediemedvetenhet inom arkivvetenskapen.

30 Internet Archives egna samlingar finns att tillgå genom en instans av Wayback Machine på adressen https://archive.org/web/web.php [Hämtad 15 maj 2014] och programvaran med samma namn, med vilken man kan ”surfa” i egna webbarkiv, är fri programvara och finns att hämta på http://archive-access.sourceforge.net/projects/wayback/ [Hämtad 15 maj 2014]

Metodbeskrivning

Detta avsnitt börjar med att beskriva hur min datainsamling *genomförts*, inklusive tekniska detaljer som möjliggör att man upprepar den. Sedan diskuterar jag det *urval* jag gjort och varför. Jag kopplar diskussionen till relevanta delar av mina *teoretiska perspektiv.* Slutligen skriver jag om hur jag har *analyserat* min data, och även de *forskningsetiska frågor* som uppstått vid insamling och bearbetning.

Genomförande

Mina data består av två olika typer av material, dels manualer, dokumentation och annat som är skrivet av människor och som jag inte kommer analysera som texter utan som jag använder för att förstå mig på Heritrix rent tekniskt. Den andra typen av material är *genererat* av Heritrix, och består av diverse olika filer som skrevs när jag samlade in min data med hjälp av spindeln, organiserade i serier med en serie för varje körning.

I en serie från Heritrix arkiv ingår de olika digitala objekt eller filer som blir kvar efter en körning, och som en konsekvens av den körningen – här kan man se exempelvis kw³ som en slags arkivbildare, med insamling (med hjälp av Heritrix) som en process i kw³s verksamhet ur vilken arkiv växer fram. E-plikten har inte samma automatiska karaktär, men genererar ett arkiv som (i idealfallet) följer vissa specifikationer. En volym ur Heritrix arkiv inkluderar

- logg-filer och en stor mängd andra filer som till viss del redovisas i *crawl-manifest.txt*, en textfil som skapas i samband med varje körning och som innehåller sökvägar till i princip alla andra loggar och filer som skapats i den körningen (cf. Sigurðsson, Stack & Ranitovic 2011:43-45),
- ett webbarkiv, som innehåller http-svar sparade seriellt i *.arc-filer*[31] i den ordning de erhållits.

Dokumentationen för Heritrix inkluderar

31 Heritrix ackumulerar varje http-svar i sin binära helhet till ett fåtal arkivfiler av formatet .arc (den senare varianten .warc innebar vissa mindre förändringar, till exempel möjligheten att utöka informationen som sparas i *header*-delen av varje http-svar). På kw³ använder man sitt eget snarlika format baserat på MIME-typer som erbjuder snarlika möjligheter, men med vissa förbättringar (enligt Allan Arvidsson i personlig korrespondens den 16 jan 2014); kw³s egna format är såpass snarlikt .warc att jag inte kommer att analysera det separat. Mer om arkivfilerna och deras format nedan i min resultatredovisning.

- manualer, främst användarmanualen författad av Sigurðsson, Stack & Ranitovic (2011) samt utvecklarmanualen författad av Halse, Mohr, Sigurðsson, Stack & Jack (2011),
- artiklar som beskriver Heritrix och som ofta är skrivna av personer involverade i utvecklingen av programvaran, exempelvis Mohr, Stack, Ranitovic, Avery & Kimpton (2004),
- Heritrix webbinterface och de olika sammanställningsmöjligheter och analysverktyg som kan användas där,
- korrespondens med kunniga bakom kw³.

E-pliktens dokumentation består av

- manualer som beskriver hur e-plikten skall följas rent tekniskt, exempelvis KB (2013b),
- lagar som styr e-plikten (främst SFS 2012:492)
- korrespondens med ansvariga för e-plikten på KB

I mitt genererande av data har jag alltså delvis följt litteraturspår, jag har funnit olika manualer och artiklar genom att söka i vetenskapliga databaser, men främst genom att följa direktioner på olika webbplatser som behandlar Heritrix, kw³ och e-plikten. Eftersom detta material används instrumentalt så oroar jag inte mig över att det ska vara *representativt* för något, utan jag har helt enkelt samlat så mycket jag känner att jag behöver för att kunna använda och förstå Heritrix tekniska sidor.

Volymen jag genererat finns förtecknad nedan,

Volym	**Seed-URI**	**Ursprung**	**Insamlat**	**Storlek (GB)**
KB-1	http://www.kb.se/	Kungliga Biblioteket	30-03-2014	6.3

Jag har valt att samla in sidor som även omfattas av e-plikten; eftersom jag inte har tillgång till e-pliktens arkiv över samma sida kan jag inte göra en *direkt* jämförelse mellan Heritrix och e-pliktens arkiv, utan de tekniska specifikationer som följer med e-plikten och hur jag tror de kommer följas får representera e-pliktens arkiv när detta blir aktuellt. Som nämnt tidigare används e-plikten som en bakgrund till Heritrix, och en direkt jämförelse mellan arkiv är inte något jag gör i min studie.

Urval

Det som styrt mitt urval är ambitionen att generera ett arkiv på ett sätt som är representativt för hur Heritrix fungerar och för hur dess arkivs form ser ut, och därför har jag helt enkelt accepterat alla de inställningar som är *default*[32], samt pekat

32 förutom att jag valt *DomainScope* som horisontmodul istället för den förhandsvalda *DecidingScope*. Scope-modulen bestämmer vilka URI-er av de som påträffat i insamlade webbsidor som ska samlas in utefter olika kriterier, exempelvis vilken domän de tillhör. DecidingScope är en nyare modul som rekommenderas i dokumentationen över de äldre horisontmodulerna (inklusive DomainScope) eftersom den är flexiblare och enklare att programmera (Sigurðsson et. al. u.å.:15), men jag valde DomainScope av tidsskäl eftersom jag fann den enklare och eftersom det var just en domänspecifik insamling, det vill säga alla sidor

spindeln mot lättillgängliga och stora webbplatser. Jag har låtit min dator samla in under flera timmar och sedan avslutat insamlingen för att spara på diskutrymme och tid. Att jag inte har ”fullständiga” inspelningar borde inte utgöra något problem för min studie, då den ju fokuserar på form snarare än innehåll; samma argument kan ges för att inte påbörja en bred insamling eller någon annan typ av insamling.

Forskningsetik

Mycket (om inte allt) av det material som finns i min volym är förstås upphovsrättsligt skyddat, och det kan tänkas att där finns diverse känsliga uppgifter av annat slag inkluderade. Eftersom mitt fokus är på arkivens *form* och inte deras *innehåll* så finns det ingen anledning att återge något av det – det vill säga webbarkivet självt – i min uppsats, så det gör jag inte, på så sätt undviker jag även alla etiska tveksamheter kring återgivning av material. Min kontakt med yttervärlden i skapandet av mitt genererade arkiv består av att jag besökt och sparat material som är fritt tillgängliga på webben – och i alla fall har jag följt gängse gällande robot-policys hos varje domän samt Heritrix inbyggda *politeness policy*, båda diskuteras mer nedan.

Vad det gäller det material som faller under *dokumentation* är det endast min personliga korrespondens med Arvid Arvidsson på KB som skulle kunna tänkas vara känslig, varför jag kommer att varsla honom om allt jag skriver med utgångspunkt i hans svar för att få hans godkännande. Korrespondensen med KB är förövrigt allmän handling. Allt annat är publicerat och fritt tillgängligt på internet.

När det gäller etiska frågor kopplat till min studie i sig, berör den ingen personligen utan endast en specifik programvara, och den innehåller ingenting som ska tas som omdömen om personer eller dylikt.

under en specifik domän såsom www.kb.se, jag var ute efter. Jag identifierade mig med min e-postadress i http-headern-fältet *from* (som skickas med varje förfrågan från spindeln till serverdatorn) (se ibid.:26).

Resultat och analys

I stil med diskursanalytiska metoder väljer jag att redovisa och analysera mina resultat löpande i samma avsnitt. Om jag delat upp redovisningen och analysen i separata avsnitt, hade redovisningen antagit karaktären av statistiksammanställningar, listor över filer med storlek, tidstämplar, listor över URIer som besökts och så vidare, vilket vore klumpigt dels eftersom många sådana listor antingen skulle vara mycket (100-tals sidor) långa eller stympade, dels eftersom denna statistik mest utgör dokumentation över *innehållet* i volymen som genererats, vilket jag ju inte är intresserad av i min studie.

Innan jag går in på volymen så kommer jag analysera och teoretisera Heritrix som programvara och medium samt dess relation till sin operativa miljö, Internet och webben. Volymen kommer sedan in för att illustrera olika avsnitt. Utöver Heritrix dokumentation och praktiskt eller tekniskt inriktade litteratur såsom Lindberg (2009) samt Croft, Metzler & Strohman (2010:33) använder jag Shannons modell (1963) för att diskutera Internet och de protokoll som utgör webben i relation till Heritrix. Beskrivningen ges på en nivå anpassad för relativt kunniga lekmän, med fotnoter som ger utökad förklaring och länkar till vidare läsning kring tekniska frågor. Vissa grundläggande delar av Heritrix tekniska funktion (exempelvis vad en URI är för något) har redan behandlats i första avsnittet.

För att knyta an till teorin disponerar jag som sagt med mediets epistemologiska konfigurationer (cf. Ernst 2013, man skulle kunna beskriva dem som begränsande eller formande faktorer i nedskrivningssystemets kanaler) som rubriknamn, vilka är *objektorientering*, *TCP/IP och HTTP, brus,* samt *sekventiell kompression*, *metadata* och slutligen beskrivs mediets *tidsperspektiv*. De följs av en *sammanfattning*.

Heritrix som programvara och medium

Introduktion

I användarmanualen introduceras vi med att ”Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler” (Sigurðsson et. al. u. å.:1), och som sådan är den i övergripande funktionalitet och algoritm lik andra spindelprogram.

Oavsett syfte följer de flesta spindlar på en abstrakt nivå samma algoritm i sin nedskrivning av webben,

1. välj URI från kölistan över URIs som schemalagts för hämtning,

2. hämta URI,
3. analysera och/eller arkivera resultatet av hämtningen,
4. lägg till alla URIs hittade som hyperlänkar och som uppfyller horisont-kriterier till kölistan över URIs att hämta
5. notera att URI är behandlad och repetera (Mohr & Stack & Ranitovic & Avery & Kimpton 2004:5).

Spindelprogram rör sig på webben på ungefär samma sätt som användarorienterade klientprogram (webbläsare), fast övermänskligt snabbt samt med övermänsklig förmåga att hantera flera sidor samtidigt – medan vi har uppe ett fåtal sidor i olika fönster, kan Heritrix hämta och behandla hundratals sidor från olika domäner per sekund. Precis som användarorienterade webbläsare[33] samlar Heritrix in sidor från webben på det sätt som föreskrivs i HTTP; användarorienterade webbläsare gömmer HTTP bakom ett interaktivt grafiskt användargränssnitt[34] medan Heritrix surfar automatiskt genom att samla och följa länkar[35].

Nedan följer *crawl-manifest.txt* från volym KB-1 i sin helhet

L+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/logs/crawl.log
L+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/logs/runtime-errors.log
L+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/logs/local-errors.log

33 Med användarorienterade webbläsare menar jag webbläsare som är byggda för att hämta och visa ett fåtal sidor för människor och tillåta navigering mellan dem, exempelvis Mozilla Firefox, se https://en.wikipedia.org/wiki/Firefox [Hämtad 12 april 2014]. Användarorienterade webbläsare framlägger resurser på Internet med många olika användarorienterade grafiskt presenterade funktioner för att underlätta mänskligt surfande, exempelvis visas hyperlänkar ofta som blåa understrukna textsnuttar, där finns ofta en bakåtknapp, man kan bokmärka sidor etc. - alla sådana funktioner är abstraktioner som använder sig av eller maskerar funktioner i protokollen som styr hur man rör sig på webben, exempelvis HTTP. Heritrix har ett grafiskt gränssnitt för att konfigurera och styra roboten, men när själva surfandet sker är det helt utan direkta grafiska representationer, och den använder sig automatiskt och mer direkt av de kommandon som HTTP medger – den är inte tänkt att användas för att surfa på webben.

34 GUI, Graphical User Interface eller grafiskt användargränssnitt, var något som introducerades för datorer på 1980-talet och som snabbt förvandlade tekniken från ett fåtal osynliga stordatorer som förprogrammerades som användes för automatiskt databearbetning för stora företag eller myndigheter, till att bli det centrala och dominanta kulturella medium för ordbehandling och annan kreativ verksamhet som vi handskas med idag (Manovich 2013:21), se https://en.wikipedia.org/wiki/GUI [Hämtad 12 april 2014]. Innan man introducerade grafiska gränssnitt programmerades datorer på förhand med hjälp av exempelvis hålkort eller genom att flytta runt kablar, output skrevs ofta ut på en pappersremsa eller på nya hålkort. Senare (exempelvis i UNIX-system och Microsoft DOS) användes textbaserade gränssnitt där kommandon skrevs in som text; det senare lever kvar idag i form av kommandotolken (i operativsystemet Microsoft Windows) eller terminalen (på UNIX-baserade operativsystem såsom Mac OSX eller Debian GNU/Linux), se https://en.wikipedia.org/wiki/Command-line_interface [Hämtad 12 april 2014]. Operativsystem är det lager programvara som sköter kommunikationen mellan hårdvara och användarorienterad programvara, se https://en.wikipedia.org/wiki/Operative_system [Hämtad 12 april 2014].

35 Ett förprogrammerat program, exempelvis Heritrix spindel eller programmet *rm* (se https://www.gnu.org/software/coreutils/manual/html_node/rm-invocation.html#rm-invocation [Hämtad 12 april 2014]) som tar bort filer, ges ett antal parametrar och arbetar sedan automatiskt enligt parametrar givna, medan interaktiva program, exempelvis en ordbehandlare eller Mozilla Firefox stannar upp och tillåter att användaren förändrar eller interagerar i programmets flöde. Gränsen mellan de två typerna är inte skarp dock, Heritrix insamling pågår automatiskt, men den kan pausas och dess instruktioner förändras vilket innebär viss interaktivitet.

L+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/logs/uri-errors.log
L+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/logs/progress-statistics.log
C+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/order.xml
C+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/seeds.txt
R+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/hosts-report.txt
R+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/mimetype-report.txt
R+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/responsecode-report.txt
R+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/seeds-report.txt
R+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/crawl-report.txt
R+ /home/simon/Programs/heritrix/heritrix-1.14.4/jobs/KB-1-20140330093712628/processors-report.txt

Varje rad innehåller en sökväg till olika filer som skapats när insamlingsjobbet KB-1 genomfördes. Några av dessa filer (märkta i fetstil ovan) behandlas i sina respektive kontexter i detta avsnittet[36]. Bokstaven i början av varje rad indikerar om filen är en loggfil (L+), en konfigurationsfil (C+) eller en rapport (R+).

Epistemologisk konfiguration: objektorientering

Etologin, läran om insekter, påverkades av och influerade under 1900-talet av cybernetiken och systemteorin, vilka i sin tur påverkade teknologiska fält att assimilera etologin som modell och inspiration för teknologiska system (Parikka 2010:129-131). Att Heritrix kallas spindel eller web crawler är alltså passande, eftersom det är just till etologisk teori som många av webbens programvaror, protokoll och strukturer mediearkeologiskt med fördel kan kopplas. Etologisk teori blev importerad för att användas i programvarudesign i form av *objektorientering*, den första epistemologiska konfigurationen som jag skriver om.

Objektorienterad design och modellering av komplexa system ”uppkom som de flesta socioteknologiska transformationerna […] inte i ett kulturellt vakuum utan vara bara ett uttryck för epistemologiska kulturella förändringar som karaktäriserade 1960-talet i stort” (Alt 2011:280). Inom datavetenskapen innebar paradigmet en inversion av rådande sätt att beskriva logiska system (exempelvis programvara) från att tidigare uppifrån-och-ner och uttömmande beskrivning till att beskriva dem som mindre beräkningsbara element (objekt) som interagerar – nerifrån-och-upp (ibid.). Objektorienteringens kulturella kontext utgjordes av en bred kategori intellektuella strömningar som brukar benämnas systemteori och cybernetik (ibid.:280-281). Vad som blev tydligt i och med datorers simulation eller modellering var fördelarna med att se program som komplexa irreducibla system som per definition är för komplexa för mänsklig förståelse eller styrning – något som politiskt hade sina rötter bland filosofer såsom Friedrich von Hayek som försvarade en klassisk liberalism som sedan utvecklades till neoliberalism. Den fria marknaden ses som en sådan där irreducibel komplexitet som inte låter sig styras av människor (Michelsen 2006:240).

Innan detta radikala epistemologiska skifte slog rot var tanken på radikalt distribuerade system, där ingen central kontroll fanns, förlagd till sinnessjukhusen och politiskt subversiv (Parikka 2010:116). Med datorns och Internets framväxt blir denna flytande form av makt eller kontroll, självorganiserande system likt bisamhällen, det definierande paradigmet i vad Terranova kallar vår nätverkskultur (2004). Samma nätverksstruktur strukturerar även vårt samtida nedskrivningssystem.

36 Om inte annat anges härstammar teknisk beskrivning från användarmanualen (Sigurðsson et. al. u. å.:12)

Objektorienterad programmering utgör ett inflytelserikt paradigm för hur programmerare tänker sig koden de skriver, och ersatte i början av 2000-talet den tidigare strukturalistiska programmeringen som ville beskriva komplexa program ovanifrån och ner genom att passa in data och algoritmer i strukturer (Crutzen & Kotkamp 2006:200). Heritrix följer en objektorienterad filosofi hela vägen från programmeringsspråket programvaran är skriven i till gränssnittet mot både mänskliga användare och Internet. Dess distribuerade anrop av hundratals olika serverdatorer samtidigt påminner om insekters svärmande beteende – man har sedan 1980-talet inom programmering använt sig av insekters levnadsmönster som modell (Parikka 2010:159) – i objektorienterad programmering byggs autonoma objekt som reagerar och kommunicera med varandra för att, likt myror eller bin, uppnå högre mål än vad varje enhet kan åstadkomma eller förnimma själv (ibid.).

Med Shannons koppling mellan information och *entropi*[37] följde ett sätt att tänka på information som lokaliserade den som en statistisk sannolikhet sprungen ur ett kommunikationssystems koder, kanaler och källor till brus, och inte bunden till någon hermeneutisk mening utan kopplad till en fysisk (teknisk) verklighet (Terranova 2004:24) – nedskrivningssystemet. Makten att ge ramarna – uppsättningarna möjliga tecken, eller entropin i systemet – ligger i stor mån hos *medieteknologier.* Nu för tiden är det maskinkod och källkod som lagrar och bearbetar diskurser istället för alfabetet, datorteknologi utför kodningen och därför praktiken i verkligheten (Kittler 2006:45). I KB-1 kan man hitta spår efter vissa av de processorer, sammankopplade som kommunicerande objekt som skickat URIer mellan varandra, i *processors-report.txt*,

```
Processors report - 201403301239
  Job being crawled:    KB-1
  Number of Processors: 13
  NOTE: Some processors may not return a report!

Processor: org.archive.crawler.fetcher.FetchHTTP
  Function:          Fetch HTTP URIs
  CrawlURIs handled: 39498
  Recovery retries:  0

Processor: org.archive.crawler.extractor.ExtractorHTTP
  Function:          Extracts URIs from HTTP response headers
  CrawlURIs handled: 39466
  Links extracted:   4040

Processor: org.archive.crawler.extractor.ExtractorHTML
```

37 Shannons sätt att använda konceptet entropi är tagen från statistisk mekanik, entropi signifierar här de antal olika möjligheter/symboler som finns i ett system av koder – information är att välja exempelvis en symbol ur en uppsättning symboler, och ju högre entropi (ju fler möjliga symboler att välja mellan, alltså ju svårare det är att förutse valet statistiskt) desto mer information ges när en symbol överförs korrekt. Entropi är ett mått på *svårigheten att förutse* eller *den ursprungliga mängden möjliga meddelanden som en kod medger*. Exempelvis skickar japanska tonåringar tomma SMS till varandra, och dessa är konceptuellt en *bit* information – informationen är att en motpart fick SMSet, istället för att inte få det (ett av två val); *meningen* bakom ett sådant SMS lär väl vara att avsändaren vill att mottagaren ska veta att hen tänker på denne, vilket kan vara nog så meningsfullt – men informationsmässigt handlar det endast om en bit, valet mellan två olika val. För en bred introduktion till entropi som även introducerar konceptet inom termodynamiken där det härstammar, se https://en.wikipedia.org/wiki/Introduction_to_entropy [Hämtad 12 april 2014].

Function: Link extraction on HTML documents
CrawlURIs handled: 13487
Links extracted: 745224

Processor: org.archive.crawler.extractor.ExtractorCSS
Function: Link extraction on Cascading Style Sheets (.css)
CrawlURIs handled: 1168
Links extracted: 18111

Processor: org.archive.crawler.extractor.ExtractorJS
Function: Link extraction on JavaScript code
CrawlURIs handled: 1306
Links extracted: 18005

Processor: org.archive.crawler.extractor.ExtractorSWF
Function: Link extraction on Shockwave Flash documents (.swf)
CrawlURIs handled: 118
Links extracted: 128

Heritrix är inte bara internt organiserat rent konceptuellt i objekt enligt objektorienterad filosofi, utan allt i programmet är objekt eftersom ett objektorienterat programspråk använts. Programmet är skrivet i Java, ett av de mest använda och spridda programmeringsspråken idag. Java består egentligen av tre komponenter: ett programmeringsspråk, *Java Virtual Machine* (JVM) samt Javaplattformen. Så här förklarar Flanagan det hela,

> The Java programming language is the language in which Java applications […] are written. When a Java program is compiled[38], it is converted into byte codes that are the portable machine language of a CPU architecture[39] known as the Java Virtual Machine [that] can be implemented directly in hardware [but most often] in the form of a software program that interprets and executes byte codes. […] The Java platform is the predefined set of Java classes[40] that

38 Kompilation är den process där kod konverteras av en kompilatorprogramvara till, ofta i flera steg, från ett högnivåspråk som Java till Java bytecode som läses av JVM och sedan konverteras till maskinkod och sedan till processorinstruktioner. Man kan säga att en dator består av ett babels torn av olika språk – på toppen finns vårt grafiska interface med knappar och en mus, och sedan går varje interaktion genom ett lager av översättning ända ner tills att hela datorns funktion är en enda serie ettor och nollor, eller skillnaden mellan en elektrisk laddnings närvaro eller frånvaro (Kittler 1997:150). För mer om kompilation, se https://en.wikipedia.org/wiki/Compiler [Hämtad 12 april 2014].

39 CPU står för Central Processing Unit och är den klump med transistorer som utför en uppsättning logiska och aritmetiska operationer enligt instruktioner (som kompilerats från källkod, skriven av människor, till maskinkod som kan läsas med viss möda av människor men som är till för att läsas av maskiner) från programvara.

40 En klass refererar till en typ av *objekt* och beskriver olika *attribut* och *metoder* som hör till den klassen – exempelvis kan man ha en klass för bilar, med attributen *färg* och *skick* samt metoderna *byt färg*. I sin kod kan man sedan *instansiera* (ung. realisera) en klass och skapa ett objekt från den klassen, exempelvis bil_1, och sedan använder vi metoden *byt färg* för att göra den röd. Sedan kan vi skapa fler bilar i andra färger etc. För en teknisk introduktion till objektorienterad programmering, se https://en.wikipedia.org/wiki/Object-oriented_programming [Hämtad 4 april 2014].

> exist on every Java installation […] available for use[41] by all programs (2005:1, fotnoterna är mina).

Heritrix har en medvetet modulärt uppbyggd arkitektur, och man använde Java för att förstärka denna modularitet. Java ansågs ha flera fördelar,

- programvara skriven i Java blir enkel att anpassa och bygga ut samt uppgradera på grund av dess inbyggda objektorienterade design,
- där finns en stor mängd programvaruutvecklare som använder sig av Java, chansen är därför större att utomstående skulle bli intresserade av att bidra till källkoden, samt mer troligt att man skulle kunna dra nytta av tidigare skriven kod i form av exempelvis klasser för att kommunicera på låg nivå[42] med webbservers. (Mohr et. al. 2004:3).

Att kunna använda sig av andra programmerares kod *utan att behöva läsa och förstå den* utgör en av de viktigaste principerna för och fördelarna med objektorienterad programmering; *inkapsling*[43] innebär att objekts funktion eller *vad de kan göra* är tillgängligt för programmerare genom ett etablerat gränssnitt. Som programmerare kan man läsa en beskrivning av gränssnittet och sedan använda objektet enligt där beskrivna regler utan att behöva bry sig om *hur* objektet uppnår de resultat det gör. Om programmerare i ett stort projekt såsom Heritrix vore tvungna att läsa och förstå varandras kod skulle ineffektiviteten öka exponentiellt, när allt de egentligen behöver veta är ”what kinds of actions the object can perform and the message format needed to request each action” (Alt 2011:296).

Objektorienterad programmering har sitt ursprung i samma laboratorium som många andra teknologiska innovationer som vi idag tar för givet, ”grafiska gränssnitt med överlappande fönster och ikoner, rastergrafik[44], färggrafik, nätverk via Ethernet[45]” (Manovich 2013:57) uppfanns alltihop på 1970-talet av Alan Kay med kollegor på Xerox Palo Alto Research Center (PARC). Det första riktigt objektorienterade programmeringsspråket *Smalltalk* utvecklades och användes här, och precis som med senare språk så skapades olika objekt som sedan kunde återanvändas i många olika program, och i denna miljö skapades även olika kreativa programvaror såsom ritprogram och ordbehandlare, datorn omdefinierades nu till att simulera inte bara alla

41 Alltså, man kan referera till – åkalla eller använda – dessa klasser i källkoden till vilket Java-program som helst, och förvänta sig att om programmet körs på någon annan dator så kommer det göra samma sak.

42 När jag skriver om hög respektive låg nivå menar jag nivå av *abstraktion* från den underliggande hårdvaran; på väldigt hög nivå ligger grafiska applikationer (såsom Mozilla Firefox), sedan talar man om högnivåspråk (såsom Java eller Ruby) och lågnivåspråk (såsom C) samt ännu lägre-nivå (Assembler). Se https://en.wikipedia.org/wiki/High-level_programming_language [Hämtad 14 april 2014]. Att kommunicera på låg nivå innebär alltså att kommunicera med hjälp av verktyg som är specialiserade och närmare detaljerna i hur internetprotokollen fungerar än när man exempelvis surfar med Firefox, där man trycker på knappar med musen etc.

43 Eng. *encapsulation*, främst i bemärkelsen att objekt är opaka, ogenomskinliga, att informationen inuti är gömd och tillgänglig endast via objektets designerade *publika* (i meningen offentliga, tillgängliga för andra objekt) metoder.

44 Eng. bitmapped display, en digital bild uppbyggd av rader av bildpunkter av specifika färger.

45 Ethernet är den teknik för lokala trådburna nätverk som dominerar på marknaden idag, antagligen eftersom den är billig, enkel att förbättra och för att de brister som finns med tekniken är väl kända; till vardags är Ethernet i princip synonymt med nätverkskort, trots att flera andra tekniker finns (Lindberg 2009:18).

andra maskiner[46] utan alla andra *medier*, alltihop samlat under ett tak och med en viss logik som följer med i varje program och som egentligen gör dem till nya medier. (ibid.). Vissa tekniker som innan datorn antingen inte fanns (såsom att automatiskt generera bilder på moln i Adobe Photoshop) eller endast kunde användas i specifika medier (att klippa ut och klistra in, exempelvis) återfinns i många medieprogram (ibid.).

Java är ett snäpp radikalare i sin objektorientering än Smalltalk i att det stoppar ett objekt mellan operativsystemet hos värddatorn och Javaprogrammet, nämligen den virtuella maskinen JVM, som översätter programmets kod till maskinkod anpassad för det specifika operativsystemet som JVM körs på. Javaprogrammet, den interpretativa miljön och operativsystemet är nu separerade ifrån varandra och kommunicerar enligt förutbestämda regler – Javas slogan lyder “Write Once, Run Anywhere” (Flanagan 2005:4). Javaprogram måste inte skrivas om för olika operativsystem, utan det enda som behöver skrivas om är JVM – denna typen av plattformsoberoende är något som eftersträvas inom LDP (Quisbert 2008:59).

Javaplattformen är en stor uppsättning verktyg i form av *klasser* och *API*er[47] som finns tillgängliga för alla program som körs på en JVM – att skriva ett program i Java är i någon mening att pussla ihop olika tillgängliga klasser för att lösa ens problem. De kan ses som en slags ordförråd för språket Java, om man ser programmeringsspråk som mänskliga språk för att styra maskiner (cf. Cramer 2006:168). Genom att använda väl etablerade klasser (exempelvis finns en uppsättning klasser i Javaplattformen som hanterar olika typer av HTTP-förfrågningar) kan utvecklarna av Heritrix hoppas på att de utvecklas och blir effektivare eftersom andra användare (som kanske använder dem i helt annorlunda projekt) bidrar till deras utveckling.

En instans av Heritrix som körs är alltså objekt skapade enligt en stor uppsättning klasser av objekt. Klasser är organiserade i hierarkier, där subklasser kan ärva egenskaper av superklasser och på så vis behandla specialfall eller på andra sätt utföra olika typer av funktioner. *org.archive.crawler.extractor.ExtractorHTML*, som används för att extrahera nya URIer ur hämtade HTML-dokument, är en subklass till en rad andra klasser, som synes i trädet nedan,

java.lang.Object
  extended by javax.management.Attribute
    extended by org.archive.crawler.settings.Type
      extended by org.archive.crawler.settings.ComplexType
        extended by org.archive.crawler.settings.ModuleType
          extended by org.archive.crawler.framework.Processor
            extended by org.archive.crawler.extractor.Extractor
              extended by **org.archive.crawler.extractor.ExtractorHTML**

(Internet Archive u.å.)

---

46 Alan Turings *Universal Computer* – Turingmaskinen – är en hypotetiskt maskin som manipulerar symboler, en i taget, på en endimensionell pappersremsa enligt en uppsättning regler, och med hjälp av dessa kan sedan vilken annan symbolmanipulerande maskin som helst simuleras. Se https://en.wikipedia.org/wiki/Universal_computer [Hämtad 14 april 2014].

47 Eng. *Application Programming Interface (API)*, en resurs (en klass eller ett objekts) ”ansikte utåt,” alltså specifikationen över hur man ska hantera ett objekt, vilka typer av värden man kan stoppa in i objektet, och vad man sedan kommer få tillbaka. Se https://en.wikipedia.org/wiki/API [Hämtad 22 april 2014].

I botten ligger klassen *java.lang.Object*, en slags ur-klass som alla andra klasser ärver från (är subklasser till).

Alan Kay förutsåg Internet när han sade att "the whole point of OOP is not to have to worry about what is inside an object. Objects made on different machines and with different languages should be able to talk to each other – and will have to in the future" (Kay, citerad i Alt 2011:296); Internet, inklusive datorer, är byggda objektorienterat från grunden, och det är just decentraliseringen i en objektorienterad design som möjliggör emergens[48] av decentraliserade nätverk. Heritrix interna struktur är formad av detta paradigm, och är i grunden uppbyggd i utbytbara *moduler* - man kallar programvaran för ett ramverk i vilket man kan bygga spindlar för specifika insamlingsjobb. Vissa kategorier av moduler (såsom scope-modulen som bestämmer horisonten för jobb) är nödvändiga, medan andra (många av de som ingår i processorkedjan som behandlar hämtade sidor för att exempelvis extrahera nya URIs ur dem) kan vara olika många till antalet och komma i olika ordning, delvis uteslutas etc.

Heritrix objektorientering sträcker sig även till sin relation till den operativa miljön – faktum är att de agenter som kontaktar webbservrar, bearbetar svaren, arkiverar dem och kontaktar frontmodulen som loggför förfarandet är självständiga dotterprocesser till Heritrix-processen som kallas ToeTreads, och vars antal bestäms automatiskt av inställningar, hårdvara och andra faktorer i omgivningen. De beter sig som insekter, objekt som arbetar tillsammans, kommunicerar med varandra och med en central överhet (de ser exempelvis till att de inte överbelastar enskilda domäner genom att samtidigt försöka hämta flera filer från dem), en design som påminner om hur exempelvis myror använder feromon-spår för att på ett decentraliserat sätt sprida information i kolonin om de närmaste källorna till föda (cf. Parikka 2010:159-162). När jag skriver om att Heritrix hämtar URIer, bearbetar dem eller dylikt, är det alltså en förenkling – egentligen sker en decentraliserad insamling av många självständiga arbetar-program som är tätt kopplade till sin operativa miljö, TCP/IP.

Epistemologisk konfiguration: TCP/IP och HTTP

I grunden sker kommunikation på Internet och andra liknande nätverk via eller enligt en familj protokoll som tillsammans kallas *Internet Protocol Suite*[49], vanligen kallat *TCP/IP* som egentligen är förkortningar på de två vanligaste protokollen som används för kommunikation på Internet, *Transmission Control Protocol* (*TCP*) och *Internet Protocol* (*IP*) (Lindberg 2009:11). TCP/IP, och direkt HTTP är Heritrix operativa miljö.

TCP/IP består av en stor mängd olika protokoll som brukar delas in i 7 hierarkiskt organiserade nivåer enligt referensmodellen *Open System Interconnection* (*OSI*) (ibid.:12),

48 Emergens i meningen att någonting uppstår som en följd av många olika entiteters interagerande på ett sätt är irreducibelt till systemets enskilda delar; återigen kopplat till insektsvärlden, där exempelvis termiters komplexa boningar uppstår ur enskildas enkla interaktioner.

49 För mer om Internet Protocol Suite, se https://en.wikipedia.org/wiki/Internet_protocol_suite [Hämtad 12 april 2014].

Nivå	OSI	Internet Protocol Suite (TCP/IP)
7	Applikation	Telnet, FTP,
6	Presentation	SMTP, HTTP, SOCKS, POP,
5	Session	IRC, HTTPS, SSH, etc.
4	Transport	TCP, UDP, DCCP etc.
3	Nätverk	IP, ICMP, ECN, etc.
2	Länk	Hårdvara (LAN-kort, Ethernet
1	Fysisk	WAN-koppling)

(ibid.:13).[50]

Man har organiserat kommunikationen i skikt av protokoll med etablerade gränssnitt emellan av samma anledning som Heritrix givits en modulär, objektorienterad design, nämligen att systemet blir enkelt att anpassa och uppdatera för olika situationer, och eftersom det blir tåligt mot fel av olika slag (ibid.).

HTTP ligger på applikationsnivån och har gått från att vara ett relativt simpelt protokoll för att överföra webbsidor (skrivna i märkspråket *HTML*[51]) till en högre grad av komplexitet och ett större användningsområde. HTTP använder sig av protokollet TCP för själva överföringen (Lindberg 2009:98), och det är på TCP-nivån som förfarandet med att etablera kontakt innan överföring sker.

Sidor på webben lagras på *webbservrar*, datorer som kör program dedikerade åt att lyssna efter och svara på förfrågningar från *klienter* (exempelvis en instans av Mozilla Firefox eller Heritrix), program som skickar förfrågningar från klientdatorer till serverprogram på webbservrar, tar emot svaren och visar eller lagrar dem i klientdatorn (Croft et. al. 2010:33). Alla sidor på webben har en URL, som specificerar vilket *schema* som används för att nå resursen, vilken *serverdomän* som har den, samt vilken resurs som efterfrågas, och ser ut såhär[52]:

`http://`	`www.dn.se/`	`ledare/`
schema	*domän*	*resurs*[53]

50 se även https://en.wikipedia.org/wiki/Template:IPstack [Hämtad 12 april 2014] där många fler protokoll finns med, länkade till respektive protokolls artikel på Wikipedia.

51 *Hyper Text Markup Language*, ett märkspråk som används för att förse innehåll på webbsidor med märken eller *taggar* som beskriver för en webbläsare hur innehållet skall visas för användaren – exempelvis innesluter man text som ska visas som en klickbar länk mellan taggarna såhär: länktext, exempelvis skulle Dagens Nyheter representeras såhär: Dagens Nyheter i en webbläsare. Se https://en.wikipedia.org/wiki/HTML [Hämtad 12 april 2014]. Detta är ett av många sätt att lägga in länka – mycket arbete läggs ner på moduler för att hitta specifika typer av länkar på sidor, exempelvis om de är gömda i dynamiska menyer etc.

52 Om inte annat anges refererar jag till Croft et. al. (2010:33-41) i den följande beskrivningen av hur HTTP fungerar.

53 Om man när man anger en URL utelämnar resurs-delen (vilket man ju ofta gör, man surfar exempelvis till http://www.dn.se/) hamnar man automatiskt på index.html eller en annan sida som ges av programservern exmpelvis baserat på vilken IP-adress som förfrågan skickas från (du hamnar automatiskt på svenska google trots att du skriver in google.com) eller på annan information om dig som sparats på servern eller på din dator i form av så kallade kakor.

När en klient ska hämta en sida specificerad med en URL kontaktar den först en *domännamnserver* (*DNS*) som översätter domännamnet till en IP-adress[54] som pekar på datorn med resursen som efterfrågas. I Heritrix fall loggas dessa kommunikationer i *crawl.log* (notera att det nedan är fråga om tre rader, men att de eftersom de är långa blivit brutna; varje rad börjar med ett litet relativt indrag, och samma gäller för alla exempel ur volym KB-1 i min uppsats),

2014-03-30T09:37:36.102Z 1 56 **dns**:sondera.kb.se LP http://sondera.kb.se/ text/dns #046 20140330093735829+14 sha1:2MUJMDF2TO2MRF25ZPK5KTXHLJOPJOS7 - -
2014-03-30T09:37:36.104Z **1** 67 dns:biblioteksstatistik.kb.se XP http://biblioteksstatistik.kb.se/?feed=rss2 text/dns #017 20140330093735933+23 sha1:PEKMZTCFCQ5OIFBCLNLYTH2S6Y6V7R3N - -
2014-03-30T09:37:36.104Z 1 57 dns:digidaily.kb.se XP http://digidaily.kb.se/?feed=rss2 **text/dns** #039 20140330093735949+13 sha1:ZQTTQB626XJXNKPTJNLJM6CYDXNCD2XI - -

Varje rad anger att sidan kontaktats via protokollet dns (fetstil första raden), att en adress erhölls (statuskod 1, fetstil andra raden), samt att svaret som mottogs var av MIME-typen text/dns (fetstil tredje raden).

En webbserver kan ha många olika program som lyssnar efter olika typer av anrop enligt olika protokoll, och för att inte behöva ta emot förfrågningar som är ställda enligt andra protokoll, lyssnar varje program till en egen *port*, helt enkelt ett nummer som designerar en specifik typ av program på en webbserver. HTTP-förfrågningar skickas enligt konvention till port 80 om inget annat specificeras i URLen. När kontakt uppnåtts med serverdatorn skickar klienten en HTTP-förfågan för att efterfråga en specifik sida; vanligast är en så kallad GET-förfrågan, som kan se ut såhär:

```
GET /ledare/stoppa-putins-terrorspel/ HTTP/1.0
```

vilket betyder ”var snäll och skicka sidan /ledare/stoppa-putins-terrorspel/ till mig, via version 1.0 av HTTP”, varpå servern först skickar tillbaka en *header*[55] följt av innehållet i filen som efterfrågats. Om man skulle skicka förfrågan via IP utan TCP som överliggande lager hade förfrågan sänts iväg utan att tidigare förbindelse upprättats, och därför hade inte klienten kunnat veta om servern just nu är mottaglig för data eller om den överhuvudtaget går att nå, vilket vore att skjuta i blindo (Lindberg 2009:17). Om det finns någonting som är fundamentalt med TCP/IP så är det att ”en applikation inte pratar direkt med IP […] TCP eller UDP[56] ska finnas emellan […] Med hjälp av IP-adresser adresserar vi själva datorn […] med hjälp av

54 Ett 32 bitar långt nummer som pekar på en dator på webben, exempelvis 89.253.61.42. Numret är uppdelat i olika segment som indikerar vilka delar av nätet man talar om, ungefär som postadresser där vi skriver personnamn, gata, stad, land. Nummer av de här typen är inte så lätta att komma ihåg och skriva in för människor, varför systemet med DNS upprättades, så att man istället kan skriva in en URL som sedan översätts till en IP-adress av domännamnservrar. Se https://en.wikipedia.org/wiki/IP-Address [Hämtad 12 april 2014].

55 En HTTP-header skickas med/innan en förfrågan eller ett svar som skickas enligt HTTP, för att definiera olika regler för den aktuella transaktionen. Den består av text i formatet namn:värde, en förfrågan till en server kan exempelvis innehålla *Accept: text/plain* och *Accept-Language: en-US*, vilket meddelar servern att man klientprogrammet vill ha ett svar i vanlig text, på (amerikansk) engelska. Se https://en.wikipedia.org/wiki/HTTP_header [Hämtad 12 april 2014]

56 *User Datagram Protocol* (*UDP*) en mindre feltålig och enklare variant av TCP som saknar delar av den senares funktioner; jag behandlar inte UDP mer i min uppsats.

TCP och UDP adresserar vi applikationer [, de] möjliggör kommunikation mellan applkationer” (Lindberg 2009:76).

TCP lägger till följande funktioner till IP (följande lista är en viss omskrivning av Lindbergs lista (2009:75)), protokollet

- gör kommunikationen *förbindelseorienterad*, det vill säga den ser till med hjälp av en handskakningsrutin[57] att mottagare och sändare är beredda för varandras kommunikation innan den sker,
- TCP tilldelar ett *portnummer* till varje program på en serverdator som lyssnar efter kommunikation, så att klientprogram kan adressera sin förfrågan till rätt programvara – exempelvis skickas frågor som vill ha svar via HTTP till webbserverprogramvara som oftast har port nr. 80,
- TCP delar upp dataflödet i lagom stora *paket;* de olika datorer som låt oss säga en webbsida ska skickas genom innan den når sin destination har olika stora *buffertar*[58], varför sidan måste styckas upp i mindre bitar. För att paket ska rekonstrueras till sidan i rätt ordning i fullständig uppsättning
- lägger TCP till dels ett *sekvensnummer* så att mottagaren kan se vilka paket som kommit fram och vilka som fattas, för att om något fattas begära
- *omsändning*, vilket TCP sköter.[59]

Adressering och att räkna med sekvensnummer är hur kommunikation på Internet fortgår utan avbrott, och för Ernst utgör den här typen av matematik en nyckel till hur vi tolkar narrativa strukturer och mer generellt hur man bör tänka sig mediasystem som arkiv som sätter ramarna för dessa tolkningar (Parikka 2013:4). Digitala datorer reducerar verkligheten till siffror vilket innebär en mediearkeologisk koppling, en ”kulturteknologisk möbiusremsa[60] mellan [Pythagoras världsåskådning i] 500-talet f. Kr. Grekland och nutiden” (Ernst 2013:157, min översättning samt min fotnot); Pythagoras trodde att världen i grunden var tal, vilket den ju är för Heritrix. ”En saker är säker. Utan TCP/IP hade vi inte haft internet” (Lindberg 2009:31).

En spindels insamling börjar med att hämta de URIs som givits som parametrar när den startats, så kallade *seed* URIs, och huruvida varje ny URI som upptäcks även hämtas bestäms sedan av insamlingens programmerade *horisont*[61], som motsvaras på en teknisk nivå av horisont-modulen. Den modul som används på denna plats avgör

57 En rutin där två parter (klientprogram och serverprogram) i början av sin kommunikation kommer överens om hur och när den ska ske, samt på vilka villkor. Se https://en.wikipedia.org/wiki/Handshake_%28computing%29 [Hämtad 12 april 2014]

58 Eng. *Buffer*, på svenska även *köminne*, en mellanstation antingen mellan en input- och en output-enhet (exempelvis ett tangentbord och en högtalare) eller mellan två bearbetningsapparaturer av något slag, där data mellanlagras när det är på väg.

59 För att illustrera hur TCP fungerar för att motverka störningar i systemet genom uppräknade åtgärder, prova att exempelvis följa en länk och sedan dra ur nätverkskabeln mitt när sidan hämtas, och sätt sedan tillbaka den igen efter några sekunder: TCP identifierar den delen av sidan som skickades men aldrig kom fram (eftersom du drog ur kabeln!) och begär automatiskt en nysändning av de paketen.

60 Om man tar en pappersremsa och vrider den ett halvt varv för att sedan tejpa ihop ändarna, får man ett möbiusband, en remsa med de (matematiskt) lustiga egenskaperna att den har endast en sida och en kant. Se https://sv.wikipedia.org/wiki/M%C3%B6biusband [Hämtad 18 april 2014].

61 På engelska *scope*, ett exempel på en horisont är de kriterier som kw³ ställer upp för att avgöra huruvida en URI är svensk eller ej, som för kw³ definierar ”svenskhet”.

för varje URI om den ska inkluderas i horisonten för insamlingen (Sigurðsson et. al. u. å.:12). I Heritrix 1.14.4 är många olika horisont-moduler valbara, som erbjuder olika sätt att på förhand definiera horisonten. BroadScope avgränsar exempelvis horisonten genom att man specificerar hur *djupt* Heritrix ska följa länkar, utan att ta hänsyn till vilken domän eller server URIn hör till, medan FilterScope låter en filtrera URIs enligt reguljära uttryck[62] applicerade på varje utvärderad URI, och DomainScope samlar in URIs om de finns under seed-domänerna (ibid.:12-14).

Grovt uttryckt skulle man kunna likna horisont-moduler med olika typer av linser eller inställningar på en filmkamera – exempelvis kan man filma i svartvitt vilket innebär att man avgränsar bort färger bortom gråskalan, och då utgör gråskalan horisonten. Shannon menade att om man ämnade lösa problemet med elektronisk kommunikation var det viktigaste att inse att problemet är att mediekanaler per definition är *brusiga*, och att uppgiften är att särskilja de signaler som skickats från en sändare från de som har sitt ursprung i en bruskälla – och detta är en definitionsfråga såväl som en teknisk fråga (Byfield 2006:127).

Skillnaden gentemot exempelvis filmkameror och andra äldre medier ligger främst i att Heritrix operativa miljö inte är passiv utan vad som samlas in beror på programvaror som körs på serverdatorer och serverdatorernas tekniska sammansättning och miljö. Heritrix operativa miljö är inte heller geografiskt lokal, utan distribuerad. Man kan se detta exempelvis på svarstider, nedan följer ett utdrag ur *crawl.log* från volym KB-1,

2014-03-30T09:37:42.999Z 404 2467 http://swepub.kb.se/robots.txt LLP http://swepub.kb.se/ text/html #045 **20140330093742955+38** sha1:ZA7TOSIA6N2FVOSJD2UXGWARMRGV72YR - -
2014-03-30T09:37:43.266Z 503 107 https://sec1.woopra.com/robots.txt XP https://sec1.woopra.com/ text/html #025 **20140330093738227+5038** sha1:N67J36CWSVSGPQLJCVMHS3EG7Q4S5VNW - -
2014-03-30T09:37:43.501Z 200 71 http://feedback.libris.kb.se/robots.txt LLP http://feedback.libris.kb.se/ text/plain #018 **20140330093743470+28** sha1:ZVK3SX6NSCJ3BT3JGIS7QJ5HHZMQONBP - -
2014-03-30T09:37:43.503Z 403 168 http://ettan.libris.kb.se/robots.txt LEP http://ettan.libris.kb.se/roller/nyheter/mediaresource/872cba04-e1ff-4d86-b517-9052a84b58b0 text/html #042 **20140330093743462+27** sha1:QBIFWUPDDAYEZ73EDPYVSI2UWEP67HAE - -

Det första numret på varje rad är tidpunkten (ner på millisekunden) som URI-förfrågan noterats i loggen, följt av vilken HTTP-svarskod[63] som mottogs, storleken på objektet som hämtades (i bytes), det hämtade objektets URI, en kod som indikerar de steg som ledde fram till att URIn hittades, URIn i vilken den hämtade URIn upptäcktes, mime-typen[64] hos det hämtade objektet, ett id-nummer som refererar till

62 På engelska *regular expressions*, ofta förkortat till *regexp*. Man använder regexp för att söka i text efter vissa mönster, och det är som ett litet programmeringsspråk som följer specifika syntaxregler (som varierar med implementation); regexpet ”[0-9]+[pm|am]” skulle ge alla textsträngar med obegränsat antal siffror följt av pm ELLER am. Sökmöjligheter enligt regexp finns i många program, exempelvis kan man ofta söka i dokumentbehandlare efter ord eller fraser med så kallade wild-cards (ofta ”*”). Se https://en.wikipedia.org/wiki/Regular_expression [Hämtad 7 april 2014]

63 Läsaren känner kanske igen 404 som ”webbsidan hittades ej” - 200 indikerar att resursen hämtades utan problem. För en lista över svarskoder och vad de betyder, se https://en.wikipedia.org/wiki/List_of_HTTP_status_codes [Hämtad 22 april 2014].

64 En sträng som indikerar vilken typ av media som resursen är angivet i formen *typ/subtyp*, exempelvis *text/plain* vilken betyder vanlig text. Se https://en.wikipedia.org/wiki/MIME_type [Hämtad 22 april 2014].

den process[65] som utförde förfrågan, och i fetstil en tidsstämpel som indikerar när hämtningen påbörjades följt av ett + och hur lång tid hämtningen tog i millisekunder, ett kondensat av innehållet i objektet, samt ett par anteckningsfält. (Sigurðsson et. al. u. å.:40).

Av intresse just nu är siffran i fetstil, där vi kan se att tiden det tar att hämta resursen (bakom +-tecknet) varierar mycket även mellan sidor som ligger under samma domän, vilket kan indikera att de härstammar från olika datorer, eller att signalen som sådan tagit olika vägar. Jämfört med en filmkamera producerar Heritrix även som vi kan se extrema mängder dokumentation över sitt förfarande och över nätverket som navigeras. Mjukvara generellt är svårt att separera från sin operativa miljö, program existerar inte som stabila och diskreta artefakter, utan som objekt som är ömsesidigt relaterade till många andra objekt enligt många olika regler, de är diskursiva objekt i Foucaults mening i att de är uttryck för bredare system av normer, processer och mönster (Yuill 2006:67) – rent tekniskt, TCP/IP.

Interaktionen med den operativa miljön styrs för Heritrix även av *robots.txt*, en speciell fil som webbadministratörer kan skapa och använda för att styra vad robotar (exempelvis GoogleBot eller Heritrix) får göra deras webbplatser. Heritrix frågar alltid efter *robots.txt* innan den börjar samla in sidor från en domän, och i filen specificeras vilka sidor som robotar bör ignorera.

Man karaktäriserar ofta insamlingar enligt vilket typ av horisont som används; breda insamlingar samlar så mycket som möjligt en gång, medan fokuserade begränsar sig till en mindre delmängd av webben (exempelvis alla sidor från en domän) och försöker istället få en så komplett insamling av den domänen som möjligt (ibid.:13). Man kan även prioritera sidor som verkar[66] handla om vissa ämnen, eller efter geografiskt läge etc.

När det kommer till webbarkivering har HTTP (till skillnad från exempelvis *File Transfer Protocol*, FTP, där man *kan* efterfråga ett manifest) en fatal svaghet i att man via protokollet inte kan efterfråga en kopia av eller ens en lista över en servers hela innehåll, utan endast kan nå resurser genom att fråga efter specifika URIer. Därför är ett spindelprogram som automatiskt upptäcker och följer länkar i hämtade dokument vad som behövs, om man nu inte har direkt tillgång till servern i fråga (Masanés 2006:23). Heritrix arkiverar enligt en metod som arbetar på klientsidan (likt vilken klient som helst) medan e-plikten försöker skapa sitt arkiv via ombud från serversidan.

Epistemologisk konfiguration: brus

Heritrix spelar in som en klient bland många. I alla signalkanaler finns som vi sett bruskällor (se Shannons diagram ovan i teoriavsnitten under rubriken *Shannon*), vi kan identifiera olika kategorier av *bruskällor* som Heritrix arbetar med,

65 Som jag beskrivit så arbetar Heritrix massivt parallellt genom att starta många underprocesser (små arbetar-program), i användarmanualen kallas de ToeTreads.

66 Detta gissas exempelvis genom att titta på länktexter, vilka typer av sidor som länkar till URIn under utvärdering etc.

1. sidor som den som designat horisonten för insamlingen inte var intresserad av men inte lyckades tekniskt avgränsa bort,
2. sidor som dagens användare av webben eller webbarkivet anser är brus, såsom skräpsidor,[67]
3. sidor som länkar till objekt vars URI inte kunnat upptäckas,
4. tekniskt brus i form av sidor i webbarkivet vars bitströmmar[68] korrumperats i något skede efter insamlingen,
5. tekniskt brus i form av fel som stöts på inom själva programvaran,

Men brus är för Shannon och Weaver inte bara något oönskat, utan någonting som är det som *definierar* kommunikation; om där inte fanns brus, så hade ju ingen kommunikation behövt ske, den *hade* inte kunnat ske, på så sätt ligger *datorvirus* för Parikka som en nyckelpunkt i nätverkskulturen, en central nod (2007:286). Brus är något som går att programmera (kring) och därför erhåller det en algoritmisk rationalitet (ibid.). Av ovanstående bruskällor är typerna 1 och 2 av sådan art att de härleds ur en användares subjektivitet, medan typ 4 är av teknisk art som tekniskt motarbetas i e-arkiv exempelvis genom att använda så kallade *kondensat*[69]. Typ 5 går att spåra exempelvis i *local-errors.log*, nedan följer en post ur volym KB-1,

```
2014-03-30T09:41:30.854Z  -2       - http://194.68.4.214/robots.txt LLRLP http://194.68.4.214/bilder/b/9789100120641.jpg no-type #048 - - - le:NoRouteToHostException@HTTP
  java.net.NoRouteToHostException: No route to host
        at java.net.PlainSocketImpl.socketConnect(Native Method)
        at java.net.AbstractPlainSocketImpl.doConnect(Unknown Source)
        at java.net.AbstractPlainSocketImpl.connectToAddress(Unknown Source)
        at java.net.AbstractPlainSocketImpl.connect(Unknown Source)
        at java.net.SocksSocketImpl.connect(Unknown Source)
        at java.net.Socket.connect(Unknown Source)
        at org.archive.crawler.fetcher.HeritrixProtocolSocketFactory.createSocket(HeritrixProtocolSocketFactory.java:131)
        at org.apache.commons.httpclient.HttpConnection.open(HttpConnection.java:708)
        at org.apache.commons.httpclient.HttpMethodDirector.executeWithRetry(HttpMethodDirector.java:387)
        at org.apache.commons.httpclient.HttpMethodDirector.executeMethod(HttpMethodDirector.java:171)
        at org.apache.commons.httpclient.HttpClient.executeMethod(HttpClient.java:397)
        at org.apache.commons.httpclient.HttpClient.executeMethod(HttpClient.java:346)
        at org.archive.crawler.fetcher.FetchHTTP.innerProcess(FetchHTTP.java:500)
        at org.archive.crawler.framework.Processor.process(Processor.java:109)
        at org.archive.crawler.framework.ToeThread.processCrawlUri(ToeThread.java:306)
        at org.archive.crawler.framework.ToeThread.run(ToeThread.java:154)
```

local-errors.log innehåller fel som stöttes på i *bearbetningskedjan* (mer om denna nedan) och som kunde hanteras av Heritrix utan att orsaka större problem med insamlingen – oftast handlar det, som i fallet ovan, om nätverksproblem av något slag. Statuskoden (-2), som har blivit satt i fetstil ovan direkt efter tidsstämpeln, indikerar att ett fel uppstod när HTTP-kontakt skulle upprättas; statuskoder mellan

67 Att jämföra med skräppost eller spam, exempelvis sidor som är designade att lura sökmotorer eller förstöra för insamlande spindlar genom att ständigt och utan slut generera stora mängder nya länkar.

68 Eng. *bitstream*, en sekvens eller tidsserie av bitar, exempelvis en sida.

69 Eng. *digest* eller *hash value*, ett sätt att koka ner (eller räkna om) en bitström (såsom en webbsida) till en kortare sträng siffror som sedan fungerar som ett fingeravtryck för den filen – om en enda bit i strömmen förändras och man räknar ut kondensaten igen så kommer den med mycket hög statistisk sannolikhet vara helt olik den första kondensaten. Se https://en.wikipedia.org/wiki/Cryptographic_hash_function [Hämtad 22 april 2014]

200 och 599 betyder samma sak som motsvarande HTTP-svarskoder, medan negativa siffror används av Heritrix för att designera andra typer av fel. Varje linje som börjar med ”at” anger namnet på den klass som objektet som stötte på problemet tillhör, med siffror som indikerar process-ID eller andra statuskoder efter vissa av dem; ursprungligen rapporterar java.net.NoRouteToHostException att man inte kunde hitta en väg till en fil som specificerades som undantag i robots.txt. (Sigurðsson et. al. u. å.:40-41)

Andra interna fel rapporteras i *runtime-errors.log.* Denna fil är tom i mitt fall, men vittnar annars om några tekniska problem stöttes på när jobbet kördes, exempelvis om datorn som Heritrix körs på fick slut på RAM-minne, eller mer troligt att någon del av programmet stött på en bugg av något slag. Fel kan även vara såpass allvarliga att de inte hinner skrivas till loggen innan Heritrix kraschar. (ibid.:41). I *uri-errors.log* finns brus i form av fel som stöts på när Heritrix försökt följa länkar, ett utdrag följer,

2014-03-30T09:38:09.322Z https://s.ytimg.com/yts/cssbin/www-embed-player-vflU2OLVa.css "Unsupported scheme: data" data:image/png
2014-03-30T09:38:09.486Z http://biblioteksstatistik.blogg.kb.se/wp-content/plugins/google-analyticator/external-tracking.min.js?ver=6.4.7.3 "http scheme specific part is too short: //" http://

Oftast beror denna typen av fel på felaktigt skrivna eller felaktigt tolkade länkar, men emellanåt stöts andra typer av fel på. Varje rad i filen har en tidsstämpel, en URI samt en kort sammanfattning av vad Heritrix tror är problemet med URIn; användarmanualen skriver att filen kan användas av experter vid felsökning, för oss illustrerar filen hur noggrant Heritrix som medium spelar in även bakgrundsbrus av den här typen.

Brus av den 3:e och 5:e typen är för mig mest intressant, eftersom de är rent tekniska och automatiska och därför avspeglar den verklighet som samlats in – de utgör en kvarleva av det förflutna på samma sätt som ett fotografi, och det är dessa två brustyper som på ett tydligt sätt skiljer Heritrix från e-plikten som medium.

Att som användare av ett webbarkiv stöta på brus av typen 3 (exempelvis i form av bilder som fattas, länkar och funktioner som inte funkar att klicka på eller att element på sidan inte verkar ligga som de ska etc.) bottnar i att Heritrix är bunden till HTTP som kanal. Objekt som hör till sidor kanske ligger dolda bakom dynamiska länkar och formulär, där Heritrix inte kan se dem. Databasdrivna webbsidor kan presentera sidor som är unika för en användare, eller kanske unika eftersom de algoritmiskt skapas baserat på klockslag eller andra faktorer. Som klient ser man endast vad en webbserver *skickar*, hur sidan *genereras* ligger helt fördolt för klienten och är helt upp till servern. När man tar ett fotografi av en rörlig scen kan man inte med hjälp av fotografiet återskapa rörelsen, men man kan åtminstone spara fotografiet och återvända till det långt efter att det den avbildar försvunnit (Roche 2006:94).

Om man här jämför med e-plikten så framgår det tydligt att en avgrund skiljer medierna åt; e-plikten är i sammanhanget ett artistiskt medium, upphovsmän ombeds att i format som de väljer (man uppmuntrar långtidsbeständiga format men ingen konvertering ska ske hos upphovsmannen (Kungliga Biblioteket 2013:1)) skicka in material som de anser utgör elektroniskt material enligt 2 § i SFS (2012:492). Paragrafen är värd att återge i sin helhet,

> 2 § I denna lag förstås med elektroniskt material en avgränsad enhet av en elektronisk upptagning med text, ljud eller bild som har ett på förhandbestämt [*sic*] innehåll som är avsett att presenteras vid varje användning (SFS 2012:492).

E-pliktens kommunikationsteknik kan kondenseras i vad som är bruskällor i e-plikten,

1. misstag eller missförstånd gällande vad som ingår i ovanstående definition av elektroniskt material i att upphovsmannen skickar in *för mycket* material,
2. när upphovsmän skickar in *för lite* material
3. när upphovsmän skickar in material med tekniska fel eller som har manipulerats/konverterats jämfört med vad som publicerades

Upphovsmän är i e-pliktens fall alltid involverade i urvalet och formatet på arkivet, och tjänstemän är involverade i andra ändan; brus av typ 1 sorteras bort, typ 2 kan antingen upptäckas eller missas, och brus av typ 3 manipuleras bort. Det ska sägas att man på KB arbetar med att utveckla många olika leveranskanaler, vissa som involverad robotinsamling (Kungliga Biblioteket 2013b:4), och en del av arbetet – men inte allt – med att rätt material levereras i rätt tid kan automatiseras[70]. E-plikten har inte som ambition att samla in webben, utan artiklar och andra publikationer som endast tillgängliggjorts online (Kungliga Biblioteket 2013c) och dessa lösa kriterium och ramar gör bruset som uppstår högst subjektivt.

Arkivet efter e-plikten kommer inte heller att dokumentera sitt eget förfarande alls såpass noggrant som Heritrix gör; leverantörsregister kommer att upprättas och arkiveras, och korrespondens med leverantörer lär också sparas, men ingen motsvarande dokumentation över *inspelningen i sig* och det brus som stöttes på kan upprättas.

Epistemologisk konfiguration: sekventiell kompression, metadata

När väl en URI bedömts tillhöra en insamlings horisont, går den igenom en serie moduler i *bearbetningskedjor* (eng. *processor chains)*, i vilka den exempelvis hämtas, genomsöks och töms på URIs att utvärdera, samt skrivs till lagringsmedium – det är upp till användaren att konfigurera egna kedjor med steg för allt som är intressant (ibid.:18-19). Den sista bearbetningskedjan städar upp efter de tidigare och skriver resultat till webbarkivet, felmeddelanden etc. till loggar, samt utvärderar alla funna URIs mot horisonten, och notifierar *fronten* (eng. *frontier*), modulen som håller koll på vilka URIs som ska hämtas, i vilken ordning, vilka som redan besökts etc., om alla nyfunna URIs samt att den hämtade URIn är hämtad (ibid.:19).

70 Mikael Johansson, teknisk expert på e-pliktavdelningen på KB skriver i privat korrespondens att metadata som bifogas inskickat material kontrolleras automatiskt med hjälp av XML-scheman, men sedan måste manuella stickprov användas för att kontrollera att det inskickade materialet motsvarar det som publicerats online [Korrespondens mottagen 5 maj 2014]. Boel Larsson som är pliktansvarig på KB nämner brus i form av exempelvis reklamsnuttar som inleder webbtv-klipp eller e-böcker med samma innehåll som den tryckta utgåvan – vissa delar av detta brus kan kontrolleras automatiskt genom att exempelvis leta efter dubletter, medan annat måste ses över manuellt i stickprov [Korrespondens mottagen 29 april 2014].

I bearbetningskedjan är modulen *ARCWriterProcessor* förvald, och skriver insamlade data till ett lagringsmedium i det logiska formatet .arc, i senare versioner av Heritrix .warc. Eftersom Heritrix laddar ner 100-tals dokument av varierande storlek och typ parallellt, och eftersom ett webbarkiv till och med i min smala insamlings fall innehåller hundratusentals hämtade objekt, fungerar det dåligt att spara varje objekt som en separat fil – det blir bland annat så många filer för operativsystemet att öppna, stänga och hålla reda på (Burner & Kahle 1996). Istället konkateneras alla insamlade objekt ihop *sekventiellt* i filer som är ungefär 100mb stora. Varje handling[71] *komprimeras*[72] individuellt när den skrivs för att möjliggöra tillgång till specificerade enskilda handlingar utan att behöva packa upp .arc-filen i sin helhet.

Formatet är självbeskrivande i att varje objekt kan identifieras och hämtas ur ett webbarkiv utan att ett externt index behövs, även om sådana index kan skapas och användas för att substantiellt öka hastigheten i sökningar. Olika typer av tillgång till ett webbarkiv stöds utan att modifikationer av webbarkivet blir nödvändiga. Varje insamlat objekt representeras i sin helhet efter en uppsättning metadatafält till varje och i varje .arc-fil separeras handlingarna med två radbrytningar. (Burner & Kahle 1996).

Nedan följer ett utdrag ur en av .arc-filerna som ingår i volym KB-1 där vi kan se metadata som hör till ett objekt,

```
http://www.dkagencies.com/images/booksimages/DK_11795LLIM_small.jpg 203.122.58.216
20140330094907 image/jpeg 2849
HTTP/1.1 200 OK
Content-Type: image/jpeg
Last-Modified: Sat, 08 Feb 2014 09:13:09 GMT
Accept-Ranges: bytes
ETag: "eaab4fcad24cf1:0"
Server: Microsoft-IIS/7.5
X-Powered-By: ASP.NET
Date: Sun, 30 Mar 2014 09:47:35 GMT
Connection: close
Content-Length: 2582
[efter här följer objektet, som är en bild]
```

Fälten efter URIn är den HTTP-header som skickades med objektet. Några av fälten är självklara, förutom några. *Accept-Ranges* inkluderat i svaret från servern indikerar att framtida förfrågningar där specifika intervaller av bytes anges är möjliga, *Etag* är ett slags ID-nummer med vilket framtida förfrågningar till servern kan referera till objektet i fråga, *Server* indikerar vilken programvara som används av servern (I detta fallet Microsoft-IIS v. 7.5), *X-Powered-By* är ett frivilligt fält där typen av teknologi som körs på sidan anges, *Connection: close* indikerar att servern efter att detta svar

71 Eng. r*ecord*, med handling menar jag här en uppsättning metadata + objektet som ingår i en .arc-fil.

72 Kompressionen sker med hjälp av *gzip*, ett fritt licensierat program som använder sig av Lempel-Ziv-kodning, en familj tekniker som är spridda i många olika implementationer och som komprimerar utan att någon information förloras. Se https://www.gnu.org/software/gzip/ för information och länkar till manual, samt https://en.wikipedia.org/wiki/Lempel-Ziv för information om algoritmen.

skickats avslutar kommunikationen, och *Content-Length* anger storleken på vad som skickats i bytes (Network Working Group 1999).

Varje .arc-fil har även ett metadatablock som beskriver .arc-filen självt, som kan se ut såhär[73]:

filedesc://IAH-20140330094907-00008-89-253-77-43.ownit.se-8080.arc 0.0.0.0 20140330094907 text/plain 1258
1 1 InternetArchive
URL IP-address Archive-date Content-type Archive-length
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<arcmetadata xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:arc="http://archive.org/arc/1.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://archive.org/arc/1.0/"
xsi:schemaLocation="http://archive.org/arc/1.0/ http://www.archive.org/arc/1.0/arc.xsd">
<arc:software>Heritrix 1.14.4 http://crawler.archive.org</arc:software>
<arc:hostname>89-253-77-43.ownit.se</arc:hostname>
<arc:ip>127.0.1.1</arc:ip>
<dcterms:isPartOf>KB-1</dcterms:isPartOf>
<dc:description>Kungliga Biblioteket</dc:description>
<arc:operator>Admin</arc:operator>
<ns0:date xmlns:ns0="http://purl.org/dc/elements/1.1/" xsi:type="dcterms:W3CDTF">2014-03-30T09:37:13+00:00</ns0:date>
<arc:http-header-user-agent>Mozilla/5.0 (compatible; heritrix/1.14.4 +http://www.no.url.se)</arc:http-header-user-agent>
<arc:http-header-from>x[74]</arc:http-header-from>
<arc:robots>classic</arc:robots>
<dc:format>ARC file version 1.1</dc:format>
<dcterms:conformsTo xsi:type="dcterms:URI">http://www.archive.org/web/researcher/ArcFileFormat.php</dcterms:conformsTo>
</arcmetadata>

Med hjälp av ovanstående metadata kan vi sluta oss till vilken programvara som skapat .arc-filen, från vilken IP-adress insamlingen skett, metadata om insamlingen som skickats till server-datorer i form av header-meddelanden innan/med förfrågningar, exempelvis <arc:http-header-user-agent> där man anger vilken (typ av) webbläsare roboten använder samt en hemsida[75] dit frågor kan ställas. <arc:robots>classic</arc:robots> anger hur robots.txt-filer följs. Denna typen av metadata, varken den rörande individuella http-svar i webbarkivet eller .arc-filerna själva, sparas inte av e-plikten; webbarkivet som efterlämnar Heritrix innehåller inte bara insamlat material utan extremt detaljerad information om medieteknologi i användning både hos organisationen som samlar in, men även om varje insamlat objekt och serverdatorn det kommer från. Mer om detta nedan i avsnittet *Diskussion och slutsatser*.

Tidsperspektiv: rumslig eller tidslig fulländning

Heritrix tidsperspektiv är komplext; webben samlas in *parallellt* och i interaktion med webbservrar över hela världen, vilket betyder att svarstider och dylikt utgör en miljö som spindeln är inseparabel från – det är svårt att finna analogier i antropomorfa perspektiv, utan de står snarare att finna insektsorienterade modeller för insamling som får sitt uttryck i objektorienterad design och programmering, som vi sett.

---

73 Blocket är skrivet i märkspråket XML som liknar HTML och där varje element av metadata omsluts av *taggar*, exempelvis <dc:description>beskrivning</dc:description>, där dc står för Dublin Core, en metadata-standard. Se https://en.wikipedia.org/wiki/Xml respektive https://en.wikipedia.org/wiki/Dublin_Core.

74 Här står min personliga e-postadress, vilken det inte finns någon anledning att ange i min uppsats, varför jag har ersatt den med ett x.

75 Eftersom jag inte har någon personlig hemsida angav jag här en sida som inte finns.

I vilken ordning som URIer hämtas kontrolleras av *frontmodulen*, varav flertalet olika beskrivs i användarmanualen men av vilka två är de som rekommenderas och vilka alltså rent tekniskt motsvarar och reglerar olika typer av tidsperspektiv för Heritrix,

1. BdbFrontier, vilket är det förvalda alternativet. BdbFrontier arrangerar köerna av URIer som ska hämtas enligt en princip om att prioritera en bred insamling (alltså, snarare sidor från många olika domäner, än sidor djupt ner i varje domäns länk-struktur) och kan konfigureras för att föredra att avsluta alla hämtningar från varje domän eller om URIer från olika domäner ska alterneras,
2. AdaptiveRevisitingFrontier, en experimentell modul som till skillnad från BdbFrontier gör återbesök till alla URIer som hittats med definierade mellanrum för att hitta innehåll som är nytt jämfört med tidigare besök, och tiden mellan besök går att automatiskt finjustera efter hur ofta URIn i fråga har förändrats tidigare under samma insamling (Sigurðsson 2005:2)

Insamlingars tidsperspektiv kan alltså separeras enligt huruvida rumslig eller tidslig fulländning eftersträvas. Antingen vill man ha en så komplett ”ögonblicksbild” (eng. *snapshot*) som möjligt av den del av webben man samlar in (alltså, så många som möjligt av alla de URIer som faller inom jobbets horisont ska samlas in), Eller så vill man ha en såpass tidsligt komplett bild som möjligt (alltså alla förändringar inom jobbets horisont ska upptäckas och samlas in, alla versioner av alla sidor ska upptäckas) (Sigurðsson 2005:2).

Att få en komplett insamling inom endera eller båda aspekterna är i princip omöjligt i praktiken, om man inte har en väldigt liten och lätthanterad horisont för sitt jobb – även om man har all hård/mjukvara i världen så är de server-datorer man samlar in från samt deras ägare begränsande i hur mycket resurser de kan och vill lägga på att svara på förfrågningar från en robot. Heritrix medger en extrem grad av definition av tidsperspektiv som även dokumenteras i *order.xml*, nedan följer ett utdrag där frontmodulens namn och inställningar återfinnes,

```
<newObject name="frontier" class="org.archive.crawler.frontier.BdbFrontier">
    <float name="delay-factor">4.0</float>
    <integer name="max-delay-ms">20000</integer>
    <integer name="min-delay-ms">2000</integer>
    <integer name="respect-crawl-delay-up-to-secs">300</integer>
    <integer name="max-retries">30</integer>
    <long name="retry-delay-seconds">900</long>
    <integer name="preference-embed-hops">1</integer>
    <integer name="total-bandwidth-usage-KB-sec">0</integer>
    <integer name="max-per-host-bandwidth-usage-KB-sec">0</integer>
    <string name="queue-assignment-policy">org.archive.crawler.frontier.HostnameQueueAssignmentPolicy</string>
    <string name="force-queue-assignment"/>
    <boolean name="pause-at-start">false</boolean>
    <boolean name="pause-at-finish">false</boolean>
    <boolean name="source-tag-seeds">false</boolean>
    <boolean name="recovery-log-enabled">true</boolean>
    <boolean name="hold-queues">true</boolean>
    <integer name="balance-replenish-amount">3000</integer>
    <integer name="error-penalty-amount">100</integer>
    <long name="queue-total-budget">-1</long>
    <string name="cost-policy">org.archive.crawler.frontier.ZeroCostAssignmentPolicy</string>
```

```
  <long name="snooze-deactivate-ms">300000</long>
  <integer name="target-ready-backlog">0</integer>
  <string name="uri-included-structure">org.archive.crawler.util.BdbUriUniqFilter</string>
  <boolean name="dump-pending-at-close">false</boolean>
</newObject>
```

order.xml är en operativ artefakt – med vissa justeringar kan man med hjälp av filen genomföra en insamling med exakt samma inställningar och modulval som en tidigare insamling; en automatisering av vad e-plikten måste använda många subjektiva intermediärer för. Order.xml är som ett frekvensnummer som ställs in på en radio, en 50 år gammal radio kan ställas in på exakt samma frekvens – och funka på samma sätt – som den gjorde för 50 år sedan, endast innehållet är som överförs är annorlunda (cf. Ernst 2013). Filen ger oss en möjlighet att komma åt medieteknologins materialitet bortom den symboliska värld som innehållet obevekligen framkallar, med den kan vi se materialitet utan att bländas av det representativa symboliska (cf. Olsen 2003:93).

Sammanfattning

Heritrix som medium bärs upp av följande epistemologiska konfigurationer,

1. *objektorientering*, ett designparadigm och en filosofi som genomsyrar Heritrix, språket programvaran är skriven i, designen av dess interna struktur såväl som dess relation till omvärlden samt dess operativa miljö,
2. *TCP/IP och HTTP*, Heritrix är tätt sammanbunden med sin operativa miljö i form av webben, vars transversala centrum utgörs av dessa protokoll,
3. *brus*, de olika typer av brus som spelas in av Heritrix är vad som utmärker mediet jämfört med andra medier och vad som låter oss identifiera det som ett tekniskt medium,
4. *sekventiell kompression*, *metadata*, är de två konfigurationer som styr Heritrix skrivande av sitt arkiv och därför mediets koppling till framtida användare.

Mediets tidsperspektiv domineras av massiv parallell bearbetning organiserad av frontmodulen samt en tidsriktning som styrs av horisontmodulen. Tidsperspektivet kopplas snarare till etologin än till antropomorfa tankemodeller.

Diskussion och slutsatser

I detta avsnitt sammanfattar jag, drar *slutsatser* och kopplar dem till mina forskningsfrågor. Jag avslutar med att reflektera över min *metod/teori* fungerat och med att föreslå *framtida forskning*.

Svar på forskningsfrågor

För tydlighets skull återger jag här min problemställning,

Hur fungerar Heritrix som arkiveringsmedium för webben, jämfört med e-plikten?

Jag operationaliserade den till följande forskningsfrågor, med underfrågor:

1. Vad karaktäriserar Heritrix som **medium** (även jämfört med e-plikten)?
   - Vad är dess epistemologiska konfiguration och specifika tidsperspektiv [time-criticality] (cf. Ernst 2013, Parikka 2010)?
   - Hur fungerar Heritrix i nedskrivningssystem 2000?
2. Hur kan Heritrix **arkiv** förstås utifrån ett mediearkeologiskt perspektiv?
   - Vad är brus och vad är information? Vad är det ett arkiv över?
   - Vad skrivs ned, jämfört med e-plikten?
3. Kan en förståelse för inspelningsmedium som mer än neutrala databärare gagna **arkivvetenskapen**?

Fråga 1 och 2 besvaras i resultatanalysen, fråga 3 i teoriavsnittet, nedan ges en kort sammanfattning av vad min studie utmynnar i för syn på Heritrix.

Heritrix och E-plikten som medier

I min resultatanalys framgår det att Heritrix fungerar som ett tekniskt inspelningsmedium. Heritrix agerar mot webben ungefär som en filmkamera agerar mot mänskliga aktörer – precis som vi kan förväxla en filmkameras blick på bioduken med en människas rent estetiskt och genom att aktivt glömma att vi tittar på just en bioduk, så agerar Heritrix till synes som vilken klientprogramvara som anropar serverdatorer på webben som helst. Men där finns flera stora skillnader, och precis som fotografiska tekniker så sparar eller ”upplever” Heritrix det *reala* i en situation, det är ”naturen själv”, i form av ljus respektive länkar – måhända skapade mer eller mindre indirekt av människor eller maskiner – som styr vad som sparas, inom ramarna för mediet och dess förprogrammering.

Både Heritrix och filmkameran medger mer eller mindre möjlighet att kontrollera och automatisera *urvalet*, man kan exempelvis peka en kamera åt ett specifikt håll och på så sätt styra urvalet, eller programmera Heritrix urval in i minsta detalj med hjälp horisontmodulen, filter av olika slag samt vid valet av seed-URIer. I någon mån är dock både kameran och Heritrix när man väl satt igång dem automatiska i sitt insamlande, de är *tekniska medier* i Kittlers och Ernst bemärkelse, snarare än artistiska medier (såsom klassiskt måleri) där urvalet i uteslutande utsträckning styrs av en artists hand.

E-plikten, däremot, går inte att jämföra med en filmkamera, utan här ligger artistiska medier närmare till hands, något som illustrerades i e-pliktens relation till brus ovan. *Varje bruskälla hos e-plikten härstammar i subjekt*. Om Heritrix, med en maskins kyliga blick, filmar en webb i rörelse och då inte får med hela rörelsen men iallafall en serie matematiskt säkra avtryck (fotografier) av vad den mottagit, så består e-plikten i att staten ber vissa delar av webben att måla den rörelse de äger och tycker faller inom ”elektroniskt material” såsom definierat ovan, och skicka in. E-pliktens relation till webben är alltså *indirekt,* den medieras genom tjänstemäns och upphovsmäns skrivarpositioner.

I nedskrivningssystem 2000 utgör e-plikten och Heritrix inte bara två olika kanaler, de skriver ner helt olika saker; e-plikten uppmuntrar i sin medieteknologiska natur artistisk litterär produktion ifrån ett segment av de som publicerar online, medan Heritrix avbildar webbens infrastruktur och vad som sänds inom en viss länkhierarkiskt definierad del av den, sett från en nod i nätverket. Det medium som är innehållet i e-plikten är det gamla mediet litteraturen, medan Heritrix innehåll är det nya tekniska mediet TCP/IP och HTTP. Detta blev även tydligt när de båda mediernas brus analyserades. Heritrix brus är direkt kopplat till/producerat av TCP/IP, webben, medan e-pliktens brus producerar ett arkiv som i bästa fall utgörs av korrespondens mellan leverantörer och tjänstemän, utan någon som helst relation till webben.

En analogi vore att se e-plikten som ett system där upphovsmän skriver av och skickar in artiklar (men inte exempelvis reklaminslag, logotyper eller annat som de anser faller utanför plikten) ur sina tidningar, medan Heritrix tar hela tidningen och även *allt annat som produceras* och sparar som det är själv, utan inblandning från upphovsmän mer än att de kan designa sina robots.txt-filer. Frågan för KB kan polemiskt uttryckas inte som ”hur ska vi bevara webben?” utan snarare som ”ska vi bevara webben, eller litteraturen från e-plikten, eller båda?”

Utvärdering av metod och teori

Jag upplevde följande fördelar med mina metod- och teorival,

- trots min litteraturs emellanåt abstrakta natur gav mina perspektiv skarpa teoretiska gränser att avgränsa längs, jag fick fram ett smalt, effektivt organiserat studieobjekt som även var praktiskt enkelt att operera och experimentera med,

- litteraturens argumenterande och politiska aspekter (såsom motståndet mot historiografiska skildringar) gör studien viktig och genomförbar,
- de var flexibla och tillät frihet i kopplingar mellan teori och empiri, vilket var oumbärligt eftersom jag hade få tidigare exempel på förfarande att luta mig mot,
- de visade sig vara utmärkt disponerade för att verkligen belysa skillnader mellan Heritrix och e-plikten,
- eftersom de behandlar skrivande (inklusive uppsatsskrivande) och narrativ kontra uppräkning (cf. Ernst 2013) etc. på ett kritiskt sett har de inte bara gett vägledning åt *innehållet* i min uppsats, utan även *formen*, vilket har varit givande och roligt.

Mitt perspektiv har saknats i arkivvetenskapen, och kommer förhoppningsvis att vara användbart i framtiden.

Framtida forskning

Under *Avgränsningar* i det inledande avsnittet av uppsatsen beskrev jag olika områden för studie som jag valde bort, och redan där nämnde jag att samma områden vore intressanta för framtida forskning. Efter att ha genomfört min studie vill jag här räkna upp några mer specifika forskningsproblem som jag tror vore givande att undersöka,

- *E-plikten i fokus*, e-plikten är såpass ny i Sverige att jag valde att använda den endast som bakgrund till min studie av Heritrix (som är ett äldre och mer väl beskrivet medium). Framtida studier av e-plikten som medium skulle effektivt kunna beskriva exempelvis hur dess struktur påverkar historiografiska möjligheter.
- *Jämlik jämförelse*, en möjligen kvantitativ studie av hur e-plikten och Heritrix arkiv ser ut jämfört med varandra, vad som fattas hos respektive etc. Man kunde med fördel fokusera på att jämföra insamlingen av samma källa i båda arkiven. Eftersom jag inte hade tillgång till arkiven, men också eftersom jag avgränsade bort den, var denna aspekt svår att inkludera i min uppsats.
- *Användarstudie*, det vore värdefullt att göra en översiktlig sammanställning över hur olika institutioner använder Heritrix, eller en djupstudie i hur användare använder och förstår arkiven som resulterar,
- *Programvara för tillgång till webbarkiv*, en (möjligen mediearkeologisk, eller kanske kvantitativt jämförande) analys av programvara som tillgängliggör webbarkiv (exempelvis Internet Archives *wayback*[76], som likt Heritrix är fri programvara) skulle passa bra tillsammans med min studie,
- *Andra protokoll*, exempelvis RSS eller FTP, och hur dessa fungerar som medier exempelvis vid insamling med hjälp av Heritrix,

Lycka till!

76 Se http://archive-access.sourceforge.net/projects/wayback/ [Hämtad 1 maj 2014]

Bibliografi

Alt, Casey. (2011). “Objects of Our Affection: how object orientation made computers a medium” In: Huhtamo, Erkki & Parikka, Jussi (ed.). (2011). *Media Archaeology: approaches, applications and implications.* Berkeley: University of California Press.

Alvesson, Mats & Sköldberg, Kaj. (2008). *Tolkning och Reflektion: vetenskapsfilosofi och kvalitativ metod.* Lund: Studentlitteratur AB

Andersson, Tommy, Fischer, Otto & Götselius, Thomas. (2012). ”Förord”. I Kittler, Friedrich. (2012). *Nedskrivningssystem 1800 • 1900*. Sverige: Glänta produktion.

Bak, Greg. (2012). ”Continuous classification: capturing dynamic relationships among information resources”. *Arch Sci* vol. 12, ss. 287-318.

Bazerman, Charles. ”The orders of documents, the orders of activity, and the orders of information”. *Arch Sci* vol. 12, ss. 377-388.

Burnell, Mats. (1995). ”Arkiven och informationsteknologien”. I Ulfsparre, Anna Christina (red.). (1995). *Arkivvetenskap*. Lund: Studentlitteratur, ss. 100-126.

Burner, Mike & Kahle, Brewster. (1996). *Arc File Format*. https://archive.org/web/researcher/ArcFileFormat.php [Hämtad 25 april 2014]

Byfield, Ted. (2006). ”Information”. I Fuller, Matthew (ed.). (2008). *Software Studies: a lexicon*. Cambridge: The MIT Press.

Cafarella, M., Chang, E., Fikes, A., Halevy, A., Hsieh, W., Lerner, A., Madhavan, J., Mutukrishnan, S. (2008). ”Data Management Projects at Google”. *SIGMOD Record*, vol. 37 mars, nr. 1.

Chun, Wendy Hui Kyong & Keenan, Thomas (ed.). (2006). *New Media, Old Media: a history and theory reader.* New York: Routledge.

Cook, Terry. (2012). ”Evidence, memory, identity, and community: four shifting archival paradigms”. *Arch Sci* vol. 13, ss. 95-120.

Cramer, Florian. (2006). ”Language”. I Fuller, Matthew (ed.). (2008). *Software Studies: a lexicon*. Cambridge: The MIT Press.

Croft, Bruce W., Metzler, Donald & Strohman, Trevor. (2010). *Search Engines: information retrieval in practice*. Upper Saddle River, New Jersey: Pearson Education, Inc.

Crutzen, Cecile & Kotkamp, Erna. (2006). ”Object Orientation”. I Fuller, Matthew (ed.). (2008). *Software Studies: a lexicon*. Cambridge: The MIT Press.

Dollar, Charles M. (1999). ”Selecting Storage Media for Long-Term Access to Digital Records”. *The Information Management Journal*, juli.

Ernst, Wolfgang. (2013). *Digital Memory and the Archive*. Minneapolis: University of Minnesota Press.

Fear, Kathleen & Donaldson, Devan Ray. (2012). ”Provenance and credibility in scientific data repositories ”. *Arch Sci*, nr. 12, ss. 319-339.

Fischer, Otto & Götselius, Thomas. (2003). ”Redaktörernas förord: Den siste litteraturvetaren”. I Kittler, Friedrich. (2003). *Maskinskrifter: essäer om medier och litteratur*. Gråbo: Anthropos.

Flanagan, David. (2005). *Java in a Nutshell: a desktop quick reference.* Sebastopol: O'Reilly.

Fuller, Matthew (ed.). (2008). *Software Studies: a lexicon*. Cambridge: The MIT Press.

Gansing, Kristoffer. (2013). *Transversal Media Practices: media archaeology, art and technological development*. Malmö: Service Point Holmbergs.

Geijer, U., Lenberg, E. & Lövblad, H. (2013). *Arkivlagen: en kommentar*. Stockholm: Norstedts juridik.

Giaretta, David. (2011). *Advanced Digital Preservation*. Berlin: Springer-Verlag

Halse, J. E., Mohr, G., Sigurðsson, K., Stack, M. & Jack, P. (u.å.). *Heritrix developer documentation.* http://crawler.archive.org/articles/developer_manual/index.html [Hämtad 5 mars 2014]

Hayles, N. Katherine. (2005). *My Mother Was a Computer: digital subjects and literary texts*. Chicago: The University of Chicago Press.

Hertz, Garnet & Parikka, Jussi. (2012). “Zombie Media: Circuit Bending Media Archaeology into an Art Method”. *Leonardo*, vol. 45, no. 5, pp. 424-430.

Huhtamo, Erkki & Parikka, Jussi (ed.). (2011). *Media Archaeology: approaches, applications and implications*. Berkeley: University of California Press.

Huvila, Isto. (2008). ”Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management ”. *Arch Sci* nr. 8, ss. 15-36.

Iacovino, Livia & Todd, Malcolm. (2007). ”The long-term preservation of identifiable personal data: a comparative archival perspective on privacy regulatory models in the European Union, Australia, Canada and the United States ”. *Arch Sci*, nr. 7, ss. 107-127.

Ilshammar, Lars (2014, kommande) "Med e- plikten tillbaka till framtiden". I Nordin, Jonas (red.). (2014, kommande). *Information som problem*. Mediehistoriskt arkiv. Stockholm: Kungl. Biblioteket.

Internet Archive. (u.å.). *Class ExtractorHTML*. http://crawler.archive.org/apidocs/org/archive/crawler/extractor/ExtractorHTML.html [Hämtad 15 maj 2014]

Johnsson, Valerie. (2011). ”Plus ça change . . .? The Salutary Tale of the Telephone and its Implications for Archival Thinking about the Digital Revolution ” *Journal of the Society of Archivists*, vol. 32, nr. 1 april, ss. 79-92.

Johnston, John. (1997). ”Friedrich Kittler: Media theory after poststructuralism”. I Kittler, Friedrich. (1997). *Literature, media, information systems: essays.* Nederländerna: G+B Arts International.

Kirschenbaum, Matthew G. (2012). *Mechanisms: new media and the forensic imagination*. Cambridge: The MIT Press.

Kittler, Friedrich. (1997). *Literature, media, information systems: essays.* Nederländerna: G+B Arts International.

Kittler, Friedrich. (2003). *Maskinskrifter: essäer om medier och litteratur*. Gråbo: Anthropos.

Kittler, Friedrich. (2006). “Code”. In: Fuller, Matthew (ed.). (2008). *Software Studies: a lexicon*. Cambridge: The MIT Press.

Kittler, Friedrich. (2010). *Optical Media:Berlin lectures 1999*. Cambridge: Polity Press.

Kittler, Friedrich. (2012). *Nedskrivningssystem 1800 • 1900*. Sverige: Glänta produktion.

Kluitenberg, Eric. (2011). On the Archaeology of Imaginary Media. I Huhtamo, Erkki & Parikka, Jussi (ed.). (2011). *Media Archaeology: approaches, applications and implications*. Berkeley: University of California Press.

Kungliga Biblioteket. (2013a). *Rekommendationer för tekniskt format i samband med e-pliktleveranser via nätverk.*

http://www.kb.se/dokument/Pliktleverans/Rekommendationer%20tekniskt%20format%20vid%20e-pliktleveranser.pdf [Hämtad 29 april 2014]

Kungliga Biblioteket. (2013b). *Introduktion till metadata i leveranser av elektroniska dokument till KB.* http://www.kb.se/namespace/digark/deliveryspecification/metadataintro/ [Hämtad 5 mars 2014]

Kungliga Biblioteket. (2013c). *E-plikt: pliktleverans av elektronisk material.* http://www.kb.se/plikt/Eplikt/ [Hämtad 29 april 2014]

Latham, Kiersten F. (2010). ”Medium Rare: Exploring Archives and their Conversion from Original to Digital: Part One: Lessons from the History of Print Media”. *LIBRES Library and Information Science Research Electronic Journal*, vol. 20, nr. 2 september.

Latham, Kiersten F. (2011). ”Medium Rare: Exploring Archives and their Conversion from Original to Digital: Part Two – The Holistic Knowledge Arsenal of Paper-based Archives”. *LIBRES Library and Information Science Research Electronic Journal*, vol. 21, nr. 1 mars.

Lindberg, Håkan. (2009). *Introduktion till IP – Internet Protocol: en guide om hur internettrafiken fungerar.* Stockholm: .SE (Stiftelsen för internetinfrastruktur).

Lomborg, Stine. (2012). ”Researching Communicative Practice: web archiving in qualitative social media research”. *Journal of Technology in Human Services*, vol. 30, s. 219-231.

Lasfargues, France & Martin, Chloé & Medjkoune, Leïla. (2012). ”Archiving before Losing Valuable Data? Development of Web Archiving in Europe”. *BFP*, vol. 36 mars, s. 118-125.

Manovich, Lev. (2001). *The Language of New Media.* Cambridge: The MIT Press.

Manovich, Lev. (2013). *Software Takes Command: extending the language of new media*. New York: Bloomsbury Academic.

Masanés, Julien. (2006). *Web Archiving*. Berlin: Springer-Verlag.

Michelsen, Anders. (2006). ”The Imaginary of the Artificial: Automata, Models, Machinics – on promiscuous modeling as precondition for poststructuralist ontology”. I Chun, Wendy Hui Kyong & Keenan, Thomas (ed.). (2006). *New Media, Old Media: a history and theory reader.* New York: Routledge.

Mohr, G., Stack, M., Ranitovic, I., Avery, D. & Kimpton, M. (2004). “An Introduction to Heritrix: an open source archival quality web crawler”. *4th International Web Archiving Workshop*, 2004.

Nerone, John. (2011). ”Genres of Journalism History”. I Robertsson, Craig (ed.). (2011). *Media History and the Archive*. Oxon: Routledge. Reproduktion av *The Communication Review*, vol. 13, nr. 1.

Network Working Group. (1999). *RFC 2616: Hypertext Transfer Protocol – HTTP/1.1*. http://tools.ietf.org/html/rfc2616 [Hämtad 26 april 2014]

Neumann, Iver B. (2003). *Mening, Materialitet, Makt: en introduktion till diskursanalys*. Lund: Studentlitteratur.

Olsen, Bjørnar. (2003). ”Material Culture After Text: Re-Membering Things”. *Norwegian Archaeological Review*, vol. 36, nr. 2.

Peters, John Durham. (2011). ”Why We Use Pencils and Other Thoughts on the Archive (An Afterword)”. I Robertsson, Craig (ed.). (2011). *Media History and the Archive*. Oxon: Routledge. Reproduktion av *The Communication Review*, vol. 13, nr. 1.

Parikka, Jussi. (2007). *Digital Contagions: a media archaeology of computer viruses*. New York: Peter Lang.

Parikka, Jussi. (2010). *Insect Media: an archaeology of animals and technology*. Minneapolis: University of Minnesota Press.

Parikka, Jussi. (2013). ”Archival Media Theory: An introduction to Wolfgang Ernst's Media Archaeology”. I Ernst, Wolfgang. (2013). *Digital Memory and the Archive*. Minneapolis: University of Minnesota Press.

Quisbert, H. (2008). *On Long-term Digital Preservation Information Systems: A Framework and Characteristics for Development*. Luleå: Universitetstryckeriet

RA-FS 1991:1. *Riksarkivets föreskrifter och allmänna råd om arkiv hos statliga myndigheter*. Stockholm: Norstedts Tryckeri AB.

RA-FS 2009:1. Riksarkivets föreskrifter och allmänna råd om elektroniska handlingar (upptagningar för automatiserad behandling). http://www3.ra.se/ra-fs/ra-fs_2009-01.pdf [Hämtad 27 april 2014]

Robertsson, Craig (ed.). (2011). *Media History and the Archive*. Oxon: Routledge. Reproduktion av *The Communication Review*, vol. 13, nr. 1.

Roche, Xavier. (2006). ”Copying Websites”. I Masanés, Julien. (2006). *Web Archiving*. Berlin: Springer-Verlag.

Rojas, Estrella. (2009). ”New figures of web textualities: from semio-technical forms toward a social approach of digital practices ”. *Arch Sci*, nr. 8, ss. 227-246.

Scannell, Paddy. (2011). ”Television and History: Questioning the Archive”. I Robertsson, Craig (ed.). (2011). *Media History and the Archive*. Oxon: Routledge. Reproduktion av *The Communication Review*, vol. 13, nr. 1.

SFS 2012:492. *Lag om pliktexemplar av elektroniskt material*. Stockholm: Utbildningsdepartementet.

Shannon, Claude E. (1948). ”A Mathematical Theory of Communication”. *The Bell System Technical Journal*, vol. 27 ss. 379–423, 623–656, juli, oktober

Terranova, Tiziana. (2004). *Network Culture : politics for the information age*. London: Pluto Press.

Weaver, Warren. (1998). ”Recent Contributions to the Mathematical Theory of Communication”. I Shannon, Claude E. & Weaver, Warren. (1998). *The Mathematical Theory of Communication*. Chicago: University of Illinois Press

Sigurðsson, K., Stack, M. & Ranitovic, I. (u.å.). *Heritrix User Manual*. http://crawler.archive.org/articles/user_manual/index.html [Hämtad 5 mars 2014]

Sigurðsson , K. (2005). *Incremental crawling with Heritrix*. http://iwaw.europarchive.org/05/papers/iwaw05-Sigurðsson.pdf [Hämtad 6 april 2014]

Snickars, P. (2011). ”Archival transitions: some digital propositions”. I: Bolton, K. & Olsson, J. (red.) (2011). *Media, Popular Culture, and the American Century*. Stockholm: KB/John Libbey Press

Sobchack, Vivian. (2011). Afterword: media archaeology and re-presencing the past. I Huhtamo, Erkki & Parikka, Jussi (red.). (2011). *Media Archaeology: approaches, applications and implications.* Berkeley: University of California Press.

Spaniol, Marc & Denev, Dimitar & Mazeika, Arturas & Weikum, Gerhard & Senellart, Pierre. (2009). “Data Quality in Web Archiving”. *WICOW'09*, April 20, 2009, Madrid.

Vismann, Cornelia. (2008). *Files: Law and Media Technology*. Stanford: Stanford University Press.

Yuill, Simon. (2006). ”Concurrent Versions System”. I Fuller, Matthew (ed.). (2008). *Software Studies: a lexicon*. Cambridge: The MIT Press.