Skapat:	*Skriven av:*	
080730	David Hansson, Lennart Stark och Gunilla Wiberg	

Digitalisering av kulturarvet
Projektrapport

Sammanfattning

Dokumentet beskriver projektarbetet för kursen Digitalisering av kulturarvet, 15 högskolepoäng vid Högskolan i Borås (distanskurs under VT 08).
Dokumentet innehåller beskrivning av projektarbetets omfattning och utförande. Observera att dokumentet kan underkastas revision under projektarbetets gång. Detta styrs med versionsnummer. Version 1.0 är första utgåvan av denna projektrapport.

Innehåll

1 Bakgrund

Projektet löper under VT-08 med start 080125 och avslut 080823. Gruppens medlemmar studerar vid Borås Högskola på distans. Gruppmedlemmar är David Hansson, Lennart Stark och Gunilla Wiberg. I gruppform kommer ett faktiskt digitaliseringsprojekt att simuleras. Handledare för kursen ”Digitalisering av kulturarvet” är Jan Buse, Mats Dahlström och Mikael Gunnarsson.

1.1 Syfte och mål

Det generella syftet med projektet är att vinna erfarenhet av det totala digitaliseringsflödet från urval till publicering.

1.2 Mål

Vårt mål med projektet är att sätta upp en miljö/system skalbart för storskalig digitalisering av kulturarvet samt utforska uppmärkning av text och finna en realistisk nivå för storskalighet.

1.3 Målsättningar

Projektets målsättningar utifrån kursdefinition är:

Att bildkällfilerna skall ha en faksimilversion (i pyramidtiff) som skapas och tillgängliggörs via webben genom användning av eRez presentationsverktyg.

Att textkällfil (TEI) skall användas för presentation av minst två olika textversioner passande målgrupp.

Projektets slutliga textfiler skall vara validerade.

Webbplatsen skall vara funktionell och användarvänlig.

Det digitala materialet (både TEI-fil och XHTML-filer) skall förses med dolda och synliga metadata. Adekvata uppgifter om det analoga originalmaterialet skall tillhandahållas, åtminstone i en TEI Header.

Källfiler skall göras tillgängliga för kursens deltagare på projektgruppens webbplats.

1.3.1 Målgruppsdefinition

Tre tänkbara målgrupper

- Barn, förskolor, grundskolor och bibliotek.
- Den akademiska världen; t ex historiker
- Allmänheten.

1.3.2 Material

Materialet är en barnbok, tryckt 1850.

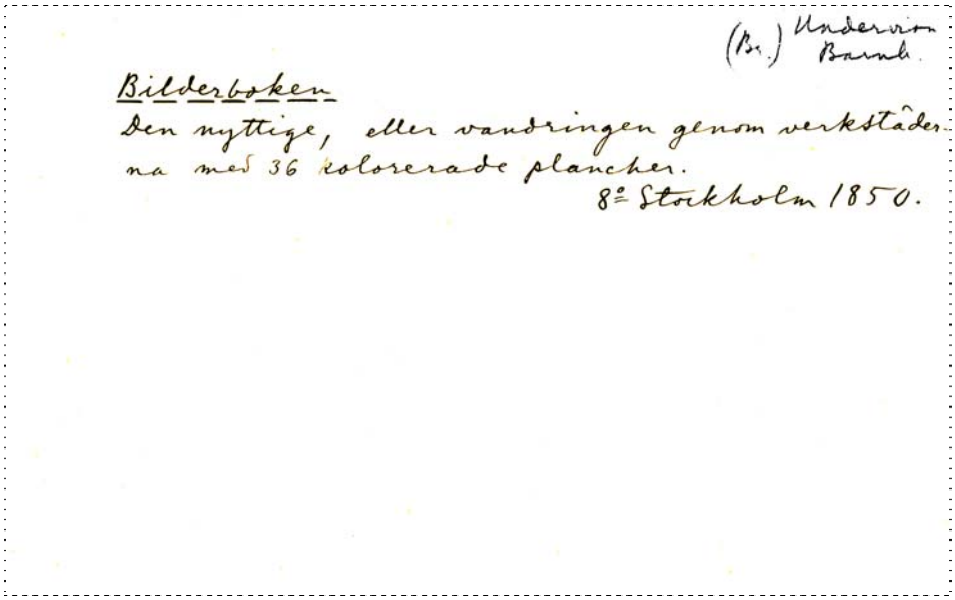
1.3.3 Beskrivning

Materialet som är valt är boken *Den nyttige bilderboken eller Vandringen genom verkstäderna*. Utgiven i Stockholm 1850 av J. C. Hedbom & Co:s förlag. Nedanstående marc-post är ett exemplar från Svenska barnboksinstitutet.

Fältnamn	Indikatorer	Metadata
000		00633cam a22001817a 4500
001		10389573
003		SE-LIBR
005		20070322094657.0
008		070322s1850 sw j 000 0 swe c
040		a Sbi
245	0 4	a Den nyttige bilderboken eller Vandringen genom verkstäderna/c med 36 kolorerade planscher
260		a Stockholm :b J. C. Hedbom,c 1850e (Stockholm :f Joh. Beckman)
300		a 16 s. [9] pl.bl. ;c 17x15 cm
500		a Litorgri, handkolorerad
510	2	a K-nr: 1065
841		5 Sbia x ab 0703224u 8 1001uu 0901128e 4
852		5 Sbib Sbih Peyron refi läsesalslån
852		5 Sbib Sbiz I Sbi marmorerat pappband

1.3.4 Placering, befintlighet, skick och läsbarhet

Nedan visas ett katalogkort (Kat-57) från Universitetsbiblioteket i Lund (skannad och retuschad). Boken är placerad i ett magasin på Universitetsbiblioteket i Lund. Skick och läsbarhet är god.


(Bx.) Undervisn. Barnb.

Bilderboken
Den nyttige, eller vandringen genom verkstäder-
na med 36 kolorerade plancher.
8:o Stockholm 1850.

Omfattningen av boken är 16 textsidor och 9 planschsidor med 36 bilder. Bokens faktiska storlek är 16x13,5 cm.

Boken innehåller en beskrivning av 36 olika yrken beskrivna av fader till hans tre barn.
Boken skick är gott och läsbarheten är tydlig så OCR-tolkning visade sig genomförbart.
Mer om bokens historiska bakgrund finns under upphovsrättsläget.

2 Projektidé

Tanken är att tillgängliggöra ett urval ur känsliga samlingar digitalt via Internet för att i största möjliga mån undvika vardagligt slitage och kunna erbjuda nuvarande och kommande generationer tillgång till materialet. För att möjliggöra detta storskaligt behövs ett "repository" som strukturerar den digitala representationen av materialet.

2.1 Produkter

Vi planerar att använda en programvara som heter Fedora (Flexible Extensible Digital Object and Repository Architecture, http://www.fedora.info/). Detta är en öppen källkod-programvara för att skapa ett digitalt av en mängd olika typer av digitala objekt. Fedora kan hantera upp till en miljon objekt.[1]

Vi kommer i mån av tid även att utforska möjligheterna för andra programvaror såsom Greenstone och ContentDM.

2.1.1 Webb

Adobe Dreamweaver CS3 används för utformning av webbpresentationen, som utformas i enlighet med 24-timmarswebben 2.0[2] Den grafiska bildvisningen sker med den kommersiella programvaran eRez och FSI Viewer. Programspråk är ColdFusion7[3]

2.1.2 Bild

Redigering av bilder sker I Photoshop CS3. Boken skannas i sin helhet

2.1.3 E-text

Boken OCR-tolkas med hjälp av en utvärderingsversion av Abbyy Fine Reader OCR 7.0 Professional Edition och märks upp av projektgruppens medlemmar.

3 Genomförandet

Inledningsvis enligt projektplanen skannades materialet i 300dpi och därefter OCR-tolkades med Abby Fine Reader. Materialet är som ett häfte i gott skick utan problem med hård bindning och kunde därför skannas med en Konica Minolta Bizhub pro C6500. OCR-tolkning gav knappt 1 procents felmarginal vilket är godtagbart. I det fall OCR-tolkningen inte varit godtagbart hade vi inte valt att ta med boken i sin helhet. Felmarginalen korrigerades av Lennart som efter det skickade texten till David. David lade sedan grundstrukturen i texten (header, och <p>) och skickade texten tillbaka till Lennart. Efter uppmärkning av yrke, råmaterial, produkt och verktyg skickades filen vidare till

[1] http://www.searchguide.se/oa/?cat=18
[2] http://www.verva.se/verksamhetsstod/webb/vl24/
[3] http://www.adobe.com/products/coldfusion/

Gunilla som under tiden förberett css-mallarna. Uppmärkning av kursivt, entiteter och normalisering gjordes av Gunilla som efter det lade till bilderna, parsade och transformerade filen med hjälp av ColdFusion som i sin tur lades i css-strukturen.

```
  <cffile action="read" charset="utf-8" file="d:\default
website\externt\apps\borasprojekt2\dnbb080630.xml" variable="theXML">
                <cffile action="read" charset="utf-8" file="d:\default
website\externt\apps\borasprojekt2\diplomatarisk.xsl" variable="theXSL">

    <cfset theXML = xmlparse(theXML)>
                    <cfset theXSL = xmlparse(theXSL)>

      <cffile action="read" charset="utf-8" file="d:\default
website\externt\apps\borasprojekt2\dnbb080630.xml" variable="theXML">
                <cffile action="read" charset="utf-8" file="d:\default
website\externt\apps\borasprojekt2\diplomatarisk.xsl" variable="theXSL">

    <cfset generatedoutput = xmlTransform(theXML,theXSL)>

                <cfset generatedoutput  = #REReplace(generatedoutput,' ,',',',"ALL")#>
                <cfset generatedoutput  = #REReplace(generatedoutput,'\t+',",'ALL")#>
                <cfset generatedoutput  = #REReplace(generatedoutput,'\n,',',',"ALL")#>
                <cfset generatedoutput  = #REReplace(generatedoutput,'\r',",'ALL")#>

<cfset generatedoutput2 = #REReplace(generatedoutput,"<\?xml",'<!--<?xml',"ALL")#>
<cfset generatedoutput3 = #REReplace(generatedoutput2,'UTF-8"\?>','UTF-8"?>--
>',"ALL")#>

                <cfoutput>
                    #generatedoutput3#
                </cfoutput>
```

Problemet här var att ColdFusion inte accepterar relativa länkar vid inläsning. Detta var ett problem som inte gick att lösa. Utöver detta kan man i xml-filen se att före de relativa länkarna för bilderna i tillägget för dtd:n finns en http: vilket också var ett krav ställt av ColdFusion.

Xlst-filerna ändrades kontinuerligt efter behov under detta arbete och även tillägg till dtd:n. Den första ändringen av xslt-filen vad det gäller normalisering gjordes av David och resterande av Gunilla. Bilderna Gunilla använde hade Lennart delat i Photoshop CS3 och sparat som tiff-filer 650x840 plockade från faksimilen (den ursprungliga inskanningen). Det visade sig dock vara svårt att få browsern att läsa tiffarna så Gunilla ändrade filerna till Jpeg 200x258 vilket fungerade med vissa tillägg i xslt- och dtd-filer. Faksimilpresentationen

gjordes med hjälp av eRez av David och placerades i css-mallarna av Gunilla. Flashen på startsidan gjordes och lades in av Gunilla.

Bakomliggande arbete med Fedora som ”repository” har gjorts av David. Utredning av upphovsrätten och kontakt med katalogisatör för katalogisering av boken i Libris och lokalt biblioteksystem Lovisa har gjorts av Gunilla. Katalogisering behövdes för att kunna namnge filerna som får sina namn bl.a. efter unikt id i biblioteksystemet Lovisa.

Slutlig validering av XHtml-filerna gjordes av David. Små ändringar i designen justerades efter XHtml-standard.

3.1 Bearbetning av textmaterial

Skanning och OCR-tolkning enligt ovan. I det fall OCR-tolkningen visat sig otillräcklig hade vi inte valt att ta med boken i sin helhet. Manuell inmatning av texten hade varit för tidskrävande.

3.1.1 TEI-uppmärkning

Uppmärkningen gjordes i olika program beroende på vilket som svarade bäst mot den tänkta uppgiften. Dessa program var jEdit, Altova XMLSpy och Dreamweaver. Valideringen är gjord i jEdit och Altova. Valideringen hade också kunnat göras via Cold Fusion. Inledningsvis definierades vad som kunde tänkas vara intressant att märka upp i ”Den nyttige bilderboken”. Eftersom boken är en redogörelse av olika yrken låg det nära till hands att märka upp yrken, verktyg, produkter och råmaterial vilket vi ansåg var intressant utifrån vår valda målgrupp. DTD:n som vi valt att använda är: http://www.adm.hb.se/~mg/dig/XMLLab/teilitex.dtd som med en del tillägg räckte till.

3.2 Faksimilåtergivning

Bildfångsten gjordes med hjälp av en Konica Minolta Bizhub pro C6500 i 300 dpi till en pdf. Utifrån denna delades materialet upp per sida och sparades som pyramidtiffar. Lite senare i projektet gjordes en ny skanning då den första saknade de tomma uppslagen. Skanningen gjordes då i 600 dpi på samma skanner och delades upp per sida och sparades som pyramidtiffar.

Filformat:	TIFF
Bredd:	3034 pixlar
Höjd:	3724 pixlar
Upplösning:	600 dpi
Komprimering:	Ingen
Färg modell:	RGB
Kanaler:	3
Bitar per kanal:	8
Mjukvara:	Adobe Photoshop CS2 Windows

Att bilderna sparas i pyramidtiffar gör att storleken ökar något på filen eftersom data läggs till. Detta optimerar renderingen. Filerna är dock till 100 procent bakåt kompertibla. Exempel på hur det underlättar renderingen ses på: http://www4.ub.lu.se/externt/apps/borasprojekt2/page1.cfm

En automatisering av ovan bit av digitaliseringen har diskuterats och anses fullt möjlig med en fast ”vagga” så bildfångsten blir exakt, Beskärning efter fasta mått, skriptning (batcha) filerna till pyramidtiffar, namngivning och skicka till server. Vi har inte automatiserat denna bit för nuvarande projekt (projektets budget täcker inte de

kostnader det innebär) men då målsättningen är att utforska möjligheterna för storskalig digitalisering har vi dock gått igenom förutsättningarna som anses mycket goda.

Pyramidtiffarna skickades upp till eRez Imaging Server där de sedan omarbetas i en FSI Viewer med manuella inställningar beroende av hur man vill att boken ska visas. En kod genereras som läggs in i html vilken i sin tur struktureras av css-mallarna. Vi har valt att presentera vårt material som en bok/häfte.

Att beställa högupplöst material har diskuterats och kommer att vara möjlig då vi dragit samman samtliga moduler i digitaliseringsprocessen. Låntagare har då möjlighet att beställa via presentation eller hemsida och betala via vår webbshop där de redan idag betalar sina utskrifter och kopior.

3.3 Transkriberad text

På: http://www4.ub.lu.se/externt/apps/borasprojekt2/page2.cfm
ses den transkriberade texten. För att se strukturen finns en ”dump” på xml- och xsl-filen att skåda på:
http://www4.ub.lu.se/externt/apps/borasprojekt2/dumpadXMLXSL.cfm
Här parsas filen genom Cold Fusion och dumpas sedan. Det kan dock ta en del tid att läsa in ”dumpen”.

3.4 Normaliserad version

På: http://www4.ub.lu.se/externt/apps/borasprojekt2/page3.cfm
ses den normaliserade texten.

3.5 Källkod

All källkod ligger på:
http://www4.ub.lu.se/externt/apps/borasprojekt2/page5.cfm

3.6 Problem

Fyra månader in i projektet förlorade vi en projektmedlem Lennart. Utöver detta arbetar resterande projektmedlemmar 100% vilket gjorde att arbetsbördan ökade under en tid och den ursprungliga planen fick revideras något. Vi har också varit tvungna att avgränsa oss något gällande ”repository”.

4 Arbetsvärdering

Huvudområde	Antal timmar (h)
Projektplanering	80
Upphovsrättsliga frågor	40
Bildbehandling	8
Skanning	1
OCR-tolkning	1
Repository implementering	
Textbehandling	100
Webbpresentation	40
Dokumentation	80

Publicering	10
Totalt	**360**

En genomsnittlig lön på 25 000 kr/mån plus arbetsgivaravgifter (år 2008, 32,42%) ger en lönekostnad på 33 105 kr/mån ca 207 kr/h. Kostnaden för personal resurser skulle då uppgå till 74 486 kr. Utöver detta tillkommer kostnader för licenser, utrustning och webbplats vilket kan beräknas till ca 10 000 kr. Total kostnad för projektet uppgår då till 85 000 kr.

Tiderna är beräknade efter viss ”tröskel” eftersom en del av teknikerna var nytt för gruppen. Skapas en storskalig digitaliseringsprocess finns erfarenheten varpå tid och kostnader minskar. Licens och materialkostnad är utslagen på olika projekt på så sätt hålls kostnaden nere.

Repository implementering har vi inte satt upp någon tid för i detta projekt. Vi anser att tiden för att sätta upp systemet ligger utanför detta digitaliseringsprojekt dock inte den tid det tar att mata in uppgifterna. Trots detta har vi valt att bortse från den tiden i arbetsvärderingen.

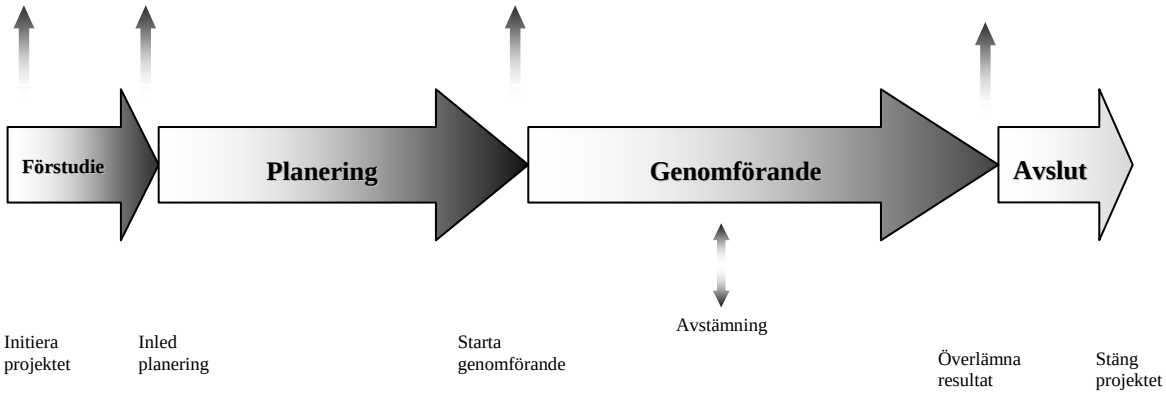
5 Projektorganisation

Samtliga i gruppen arbetar inom statlig regi och har erfarenhet av hierarki. I detta projekt arbetar vi med en platt organisation där alla har ansvarsområden och på så sätt bitvis agerar projektledare. Samtliga gruppmedlemmar kommer att vara delaktiga i alla moment oberoende av ansvarsområde. Dokumentationen görs löpande genom projektet av samtliga medlemmar. Detta sker genom versionshantering via google groups.

5.1.1 Kommunikation

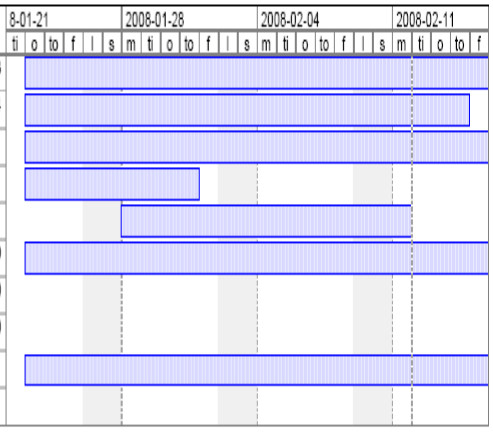
Kommunikation mellan gruppmedlemmar hålls kontinuerligt och informellt. Kontakt via e-post och möten via webb tas en gång i månaden och vid behov. Gruppen har via google groups en samlingsplats för all gemensam information och dokumentation.

5.1.2 Tidplan



5.1.3 Aktiviteter

ID	Aktivitet	Varaktighet	Start	Slut
1	Projektplanering	29 dagar?	on 08-01-23	må 08-03-03
2	Upphovsrättsliga frågor	17 dagar?	on 08-01-23	to 08-02-14
3	Bildbehandling	49 dagar?	on 08-01-23	må 08-03-31
4	Skanning	7 dagar?	on 08-01-23	to 08-01-31
5	OCR-tolkning	11 dagar?	må 08-01-28	må 08-02-11
6	Repository implementering	28 dagar?	on 08-01-23	fr 08-02-29
7	Textbehandling	60 dagar?	må 08-03-10	fr 08-05-30
8	Webbpresentation	96 dagar?	må 08-02-18	må 08-06-30
9	Dokumentering, löpande	138 dagar?	on 08-01-23	fr 08-08-01
10	Publicering	2 dagar?	on 08-07-30	to 08-07-31



5.1.4 Ansvarsplan

VEM	VAD
David Hansson	Fedora, XML, TEI, dokumentation
Lennart Stark	Bildbehandling, OCR, XML, TEI, dokumentation
Gunilla Wiberg	Webbpresentation, Flash, XML, TEI, dokumentation

5.2 Riskanalys

Händelse som kan påverka projektet negativt	Sannolikhet	Konsekvens	VÄGNING
Bortfall av gruppmedlem	1	8	8
Uppmärkning tar längre tid än beräknat	3	4	12
Repository mer tidskrävande än beräknat	3	8	24
Hårdvaru- resp mjukvaruproblem	1	9	9
Materialet håller ej	1	5	5

Förebyggande åtgärd	Åtgärd	Ansvarig	Deadline
Test av textuppmärkning på tidigt stadium	Begränsad del av boken blir uppmärkt	David	Siste mars
Repository implementeras tidigt	Samtliga involveras	David	Siste februari
Hårdvaru- resp mjukvaruproblem	Kontakter med leverantörer	Gunilla och David	löpande
Materialet håller ej	Val av nytt material	Lennart	Siste februari

6 Upphovsrättsläget

Upphovsmannen till denna barnbok är okänd. Även upphovsmannen till de handkolorerade planscherna är okänd. Eftersom boken är tryckt och utgiven 1850 är upphovsrättsläget gynnsamt. Skyddstiden för verk där författaren avlidit före 1926 är 50 år och 70 år för de verk där författaren avlidit efter 1926. Troligtvis har författaren avlidit innan 1926 men även om så inte var fallet bör skyddet ha upphört kring 2000-talet. Vi har försökt spåra upphovsmannen genom tryckeriet men då både tryckeri och förlag har upphört orsakade detta problem.

Eftersom Universitetsbiblioteket i Lund tar emot pliktleveranser från alla svenska tryckerier har vi försökt spåra tidpunkten för tryckeriets upphörande eller eventuellt övertagande/uppköp. Det visade sig vara en återvändsgränd gällande förlaget som gick i konkurs 1853. Beckmans tryckeri som 1850 tryckte boken var under en tid ett av huvudstadens förnämsta tryckerier. År 1968 förvärvades Beckmans av Svenska Repro AB, vilket 1974 uppgick i Liber AB. Vi tog därefter kontakt med Liber AB som inte har några som helst möjligheter att ta fram information om bokens upphovsman och anser sig inte ha några krav eller synpunkter på att boken blir tillgänglig via Internet.

Därefter togs kontakt med Per S Ridderstad som är docent i litteraturvetenskap och professor i bok- och bibliotekshistoria inom Lunds universitet. Inte heller han kunde hjälpa oss angående den svenske författaren och/eller illustratören men han tipsade om en tysk bok med liknande titel. ”Der technologische Jugendfreund oder Unterhaltende Wanderungen in die Werkstätte der Künstler und Handwerker”, författad av Bernhard Heinrich Blasche, fyra delar 1804-08. P. Ridderstads kommentar till detta var: ”De flesta böcker för yngre läsare

kring 1800-talets mitt hade ju tysk bakgrund. Man kan ju undra om inte både text och bilder i det här trevliga lilla häftet hämtat inspiration från en tysk förlaga, kanske en komprimering".

Efterforskningar via nätet visade att tre av dessa fyra delar fanns på Deutsches Museum – Bibliothek. Kontakt togs och efter viss övertalning skickade de några utdrag från de olika delarna. Ingen i gruppen var tillräckligt slängd i tyska eller kunde tyda skriften (frakturstil) tillfredställande för att göra jämförelsen om det kunde röra sig om en komprimering av "Der technologische Jugendfreund oder Unterhaltende Wanderungen in die Werkstätte der Künstler und Handwerker". Kontakt upprättades med en tysktalande bibliotekarie med erfarenhet av åldrat material. Denne kunde dock inte med säkerhet säga om det rörde sig om en komprimering av den tyska boken men likheterna var stora. Dock kunde vi med säkerhet säga att det inte var samma illustratör.

Göte Klingberg och Inger Bratts "Barnböcker utgivna i Sverige 1840-89" gav inte heller mer information än vad vi redan hade. Det är möjligt att om man kontrollerat förlagets/tryckeriets leverantörer på den tiden hade det kanske dykt upp ytterligare ledtrådar värda att följa men detta har vi inte gjort. Vad det gäller bilder under denna tid var det mer eller mindre fritt att använda de bilder man ansåg var passande till texten.

Det är fortfarande möjligt att boken är inspirerad av en tysk förlaga eller t.ex. en dansk som i sin tur inspirerats av en tysk förlaga, men vi bedömer att våra efterforskningar är tillräckliga och får tråkigt nog fortfarande säga att upphovsmannen är okänd. Den upphovsrättliga bedömningen av materialet är att det är fritt att tillgängliggöra på Internet.

7 Uppnådd målsättning

Målsättningen utifrån kursdefinition anser vi uppnådda. Egna uppsatta mål för projektet var att sätta upp en miljö/system skalbart för storskalig digitalisering av kulturarvet samt utforska uppmärkning av text och finna en realistisk nivå för storskalighet. Vi har dock inte satt upp miljön men utforskat möjligheterna och funnit en realistisk nivå att lägga TEI-uppmärkningen på. Att vi inte fullt ut satt upp miljön beror dels på att vi i detta projekt inte har budgeten och dels har tiden inte räckt till. Det generella syftet med projektet var att vinna erfarenhet av det totala digitaliseringsflödet från urval till publicering vilket vi anser uppnått.

7.1 Vidareutveckling

- Webbplatsen: komplettera så språk och menyer ligger i en XML-fil och byts dynamiskt
- Ytterligare TEI-uppmärkning; fler kategorier
- Test mot fler browser än vad vi i detta fall använt: Internet Explorer
- Titta på möjligheten att bygga upp en modul för TEI-användning i digitalisering av handskrifter