|  |
| --- |
| Rapport teoretisk uppgift |

Konsolidering i Docmarker

*Konsolidering av myndighetsförfattningar*

|  |
| --- |
| *Författare:* Tommy Karlsson  *Handledare:* Mikael Kjellson  *Examinator:* Johan Leitet  *Termin:* VT13  *Ämne:* Datavetenskap, självständigt arbete  *Nivå:* Högskoleexamen  *Kurskod:* 1DV42e |



Abstrakt

Myndighetförfattningar ges ut som en grundförfattning och flera förändringsförfattningar. För att förenkla läsbarheten ges även konsoliderade författningar ut där samtliga förändringar finns med. Arbetet med att skapa en konsolidering är tidskrävande. Syftet med detta arbete är att undersöka möjligheten att helt eller delvis automatisera processen att konsolidera en författning. Detta görs genom att skriva ett verktyg som använder sig av metadata i dokumentens Xml-struktur för att bestämma hur en text ska behandlas.

I de tester som genomfördes inom ramen för detta arbete halverades tiden det tog att skapa en konsolidering.

Även om detta arbete inte kom hela vägen till en helt automatiserad konsolidering kan slutsatsen dras att det finns ett behov av verktyget ifråga. Dessutom visar arbetet att författningar skapade i Docmarker har den nödvändiga metadata som krävs för att automatisera konsolideringsprocessen.

Innehåll

1 Inledning 1

1.1 Bakgrund 1

1.2 Syfte 1

1.3 Tidigare forskning 1

1.4 Frågeställning 1

2 Metod 2

2.1 Delar som berörs av arbetet är 2

2.2 Funktionalitet som måste finnas inom ramen för arbetet 3

2.3 Genomförande 4

2.4 Metod-diskussion 5

3 Resultat 6

4 Diskussion 6

5 Referenser I

# Inledning

I Domstolsverkets dokument introduktion till rättsinformationssystemet [1] ges myndigheter riktlinjer för att ge ut författningar som standardiserade Xml-dokument. Detta ger en lättare hantering av rättsliga dokument. Dessutom är det en förutsättning för att ge enkel tillgång till dessa dokument via internet.

Denna rapport kommer att undersöka möjligheten att per automatik skapa en sammanställd myndighetsförfattning.

## Bakgrund

När myndigheter ger ut förfrattningar sker detta i form av en grundförfattning. När ändringar till författningen ska göras ges det ut en ändringsförfattning, som bara tar upp de delar som ska ändras. En författning kan ha flera ändringsförfattningar som tar upp olika förändringar. I dagsläget är det ett stort klippa-och-klistra-jobb att sätta ihop en gällande sammanställd författning ur alla förändringar som förekommit genom åren. Därför har flera myndigheter visat intresse av ett verktyg som automatiserar processen att skapa en sammanställd utgåva.

Sigma Kudos har utvecklat en programvara som heter DocMarker, denna används av ett antal myndigheter i landet för att standardisera formateringen av deras författningar. Författningar ges ut i form av Xml-dokument som ett standardiserat format. Namn och attribut på element i Xml-strukturen som DocMarker lägger till, till exempel stycke och ändringsmarkeringar, gör att en automatiseringsprocess kan vara genomförbar.

## 1.2 Syfte

Rapporten kommer att undersöka möjligheten att per automatik skapa en sammanställd myndighetsförfattning eller underlätta skapandet av en konsoliderad författning.

## 1.3 Tidigare forskning

Att ge ut lagtexter som standardiserade Xml-dokument är relativt nytt. Då det är ett steg i ledet att släppa lagtexter fria lättillgängligt visa internet. I Data models for version management of legislative documents [2] har man undersökt vilka datamodeller som kan tänkas användas vid versionshantering av konsoliderade lagtexter, denna artikel har dock inte undersökt hur svenska lagtexter ser ut. Det finns en italiensk studie, NLP-based Metadata Extraction for Legal Text Consolidation [3], som behandlar hur man ska kunna konsolidera lagtexter. De extraherar metadata ifrån författningstexterna för att bestämma hur en text ska behandlas. Tack vare att de dokument som ska undersökas i detta arbete är skapade med DocMarker finns metadata redan markerat i Xml-strukturen.

## 1.4 Frågeställning

Går det att automatisera sammanställning av myndigheters föreskriftsdokument med så stor säkerhet att det ger en fördel över att sammanställa dokumentet manuellt?

# Metod

För att kunna skapa en konsoliderad författning måste först en avgränsning göras av vilka delar i författningarna som kommer att beröras inom ramen för arbetet. Nästa steg är att sätta upp regler för hur olika typer av texter ska behandlas och märkas upp i en konsoliderad författning. För att bestämma vilken text som ska behandlas och hur den ska behandlas kommer metadata i Xml-strukturen att användas.

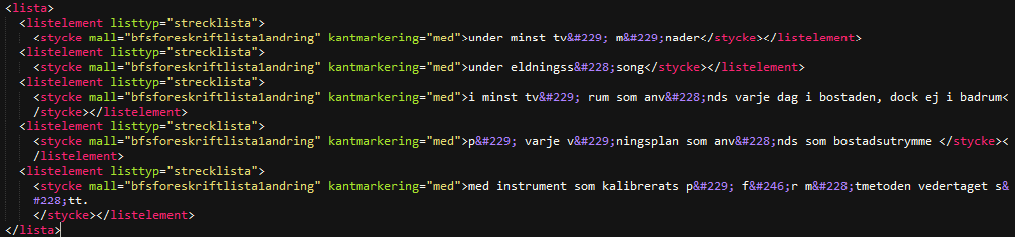
## Delar som berörs av arbetet är

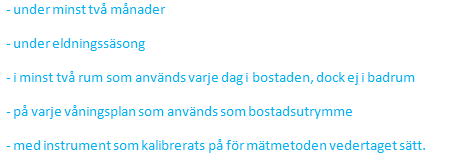
**Paragrafer**

Nya paragrafer som läggs till i en författning ska klippas in i den konsoliderade versionen.

**Författningsstycken**

Stycken som är förändringsmarkerade ska kontrolleras om de är helt nya stycken, eller förändrade versioner av befintliga stycken. Verktyget ska även i viss mån kunna identifiera stycken som har blivit borttagna ur texten.

*En ändrad text, så som den ser ut i Xml-strukturen. Att texten är ändrad ses på attributet kantmarkering.*

**

*I detta fallet var texten ändars och ska markeras med blå färg i Word-dokumentet.*

**Rubriker**

Ska behandlas som stycken och dessutom markeras att de är rubriker för att kunna formatera dem korrekt i slutdokumentet.

**Ingresser**

Samtliga ingresser i alla dokument ska identifieras och finnas med först i det konsoliderade dokumentet.

**Ikraftträdande-stycken**

Ska behandlas på samma sätt som ingresser men ligga sist i slutdokumentet.

## Funktionalitet som måste finnas inom ramen för arbetet

**Kontroll av metadata**

En funktion som söker efter specifik metadata i Xml-strukturen, för att hitta de delar i ett dokument som ska behandlas vidare av verktyget. Denna funktion ska även bestämma hur dessa delar ska behandlas.

**Textjämförelse**

Verktyget kommer att innehålla en funktion för att att jämföra texter, för att kunna bedöma om ett stycke är nytt eller förändrat.

**Dokumentjämförelse**

Det måste finnas en funktion som markerar och jämför stycken i två dokument för att kunna avgöra vilka delar i ett förändringsdokument som ska tas med i den konsoliderade författningen.

**Xml generering**

Verktyget måste kunna generera ett konsoliderat Xml-dokument utifrån de indokument som anges. Här ska alla förändringar finnas med och de beslut som verktyget har gjort i sin konsolidering ska redovisas.

**Generering av Word-dokument**

Ett Word-dokument ska genereras där förändringar och beslut som verktyget har gjort ska redovisas med färgkodningar. Detta för att det ska vara enkelt att få en uppfattning om hur väl verktyget fungerar. Färgkodningarna ska även fungera som stöd till en slutlig konsolidering.

## Genomförande

Konsoliderings verktyget kommer att skrivas i språket C# med microsofts .NET ramverk i utvecklingsmiljön Visual Studio 2010. Denna utvecklingsmiljö har valts för att det finns funktioner¹ som underlättar inläsning och behandling av Xml, samt generering av Xml och word-dokument.

För att få en fungerande textjämförelse måste först samtliga stycken markeras med i vilken paragraf stycket finns, detta gör att förändringar endast kontrolleras mot en paragraf istället för ett helt dokument. I förändringsförfattningen är ändrade eller nya stycken särskilt markerade. Om ett stycke är markerat jämförs detta med den aktuella paragrafen. Textjämförelsen bygger på en algoritm som heter Levenshtein avstånd[4]. Den fungerar genom att beräkna hur många förändringar som måste göras för att två textsträngar ska vara identiska.

Textjämförelsen används även för att kontrollera om stycken som finns i originaldokumentet saknas i ett förändringsdokument. För att inte markera text som avsiktligt har utelämnats ur en paragraf måste text som har tagits bort finnas i en sekvens med text där stycke före och efter den borttagna texten kan matchas.

Rubriker kontrolleras på samma sätt som stycken, men matchas endast mot andra rubriker.

Ingresser och ikraftträdande-stycken kan identifieras genom att de ligger utanför vanliga paragrafer och kan därigenom klippas in på respektive plats i det konsoliderade dokumentet.

När samtliga förändringsdokument har jämförts med grundförfattningen har en lista skapats där alla förändringar finns med. Denna lista skickas vidare till en klass som med hjälp av grundförfattningen och förändringslistan skapar ett konsoliderat Xml-dokument. Detta Xml-dokument används sedan för att generera ett word-dokument som utifrån markeringar i Xml-strukturen blir formaterat efter de förändringar som har skett.

Två konsoliderade dokument kommer skapas utifrån en dokumentmall för att få korrekt formattering. Det ena kommer skapas helt manuellt utifrån de indokument som en författningsserie består av. Det andra skapas ifrån det konsoliderade dokument som verktyget har skapat. En jämförelse av hur lång tid de olika konsolideringarna tar ska ge svar på om automatiseringen har lyckats.

1Linq-to-Xml <http://msdn.microsoft.com/en-us/library/bb387098.aspx>

 Microsoft Office Interop Word  <http://msdn.microsoft.com/en-us/library/microsoft.office.interop.word(v=office.14).aspx>

## Metod-diskussion

Metoden som används för att bestämma hur väl verktyget fungerar påminner till stor del hur man hade valt att göra i NLP-based Metadata Extraction for Legal Text Consolidation [3]. En risk som finns är att verktyget utifrån textjämförelsen tar beslut om ett stycke är nytt eller förändrat, det kan vara svårt att hitta en bra gräns hur lika två texter behöver vara för att i grunden vara samma stycke.

Eftersom att att samma person kommer att skapa samtliga konsolideringar kommer tidsjämförelserna vara en ungefärlig uppskattning av hur väl verktyget fungerar och inte någon säkerställd statistik. Den konsolidering som skapas med hjälp av verktyget kommer alltid att skapas först för att tiderna inte ska bli kortare på grund av att den som konsoliderar känner till dokumenten. För att i framtiden få en bättre analys av verktyget borde det användas av flera personer som jobbar med konsolideringar, på en större mängd dokument. Men den undersökning som görs fungerar som underlag till ett beslut om metoden är värd att vidareutveckla.

# Resultat

Att komma fram till en helt automatiskt skapad konsolidering är inte rimligt inom ramen för detta arbete. Verktyget kommer inte att skapa en slutlig konsolidering och den kommer inte heller att ta hårda beslut, som att ta bort text, denna kommer endast att markeras som borttagen. Verktyget kommer istället att generera ett sammanställt dokument som kan användas som grund till en färdig konsolidering.

Att skapa en konsoliderad författning, med verktygets sammanställda dokument som grund, halverade tiden som det tog att skapa samma konsolidering helt manuellt genom att gå igenom hela författningsserien för att hitta samtliga förändringar.

# Diskussion

Det finns många faktorer som påverkar vilken av metoderna som är bäst att utveckla, tidsvinst kontra utvecklingskostnad, osäkerhet i slutdokumentet beroende på specialfall i författningstexterna och hur exakt en konsolidering måste vara, både i fråga om formatering och vilka delar som rent juridiskt måste finnas med, för att den ska godtagbar att publicera. Men det finns alla möjligheter att med hjälp av den metadata som Docmarker lägger i Xml-strukturen minska arbetsinsatsen som krävs för att konsolidera författningsdokument.

Om man istället för att som i detta arbete leta efter förändrade texter utgår ifrån sekvenser av texter som man vet är identiska skulle man med större säkerhet kunna bestämma vart en förändring i författningen har skett. Om man kombinerar detta angreppsätt med en kraftfullare och mer flexibel textjämförelse, borde man kunna skapa ett konsoliderat dokument som har mindre felmarginal än den mänskliga faktorn.

# Referenser

[1] Introduktion till rättsinformationssystemet, Domstolsverket, 14 Februari 2013. [Online]

Tillgänglig: <http://dev.lagrummet.se/dokumentation/introduktion/intro-beslutsfattare.pdf> [Hämtad 2013-05-02]

[2] Data models for version management of legislative documents, Journal of information science, February 13, 2013. [Online]

Tillgänglig: <http://jis.sagepub.com.proxy.lnu.se/content/early/2013/02/18/0165551512473723.full> [Hämtad 2013-04-24]

[3] NLP-based Metadata Extraction for Legal Text Consolidation, ACM New York 2009. [Online]

Tillgänglig: <http://delivery.acm.org.proxy.lnu.se/10.1145/1570000/1568240/p40-spinosa.pdf?ip=194.47.65.106&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF10545DFF9339251C007439F84D856B744&CFID=217745441&CFTOKEN=81043658&__acm__=1368780175_0dcd994350813a0363f60ae5d0d83910> [Hämtad

2013-05-17]

[4] Wikipedias artikel om Levenshtein-avstånd. [Online]

Tillgänglig: <http://en.wikipedia.org/wiki/Levenshtein_distance> [Hämtad 2013-05-10]