H.C. Andersens Eventyr og Historier.
Den digitale manuskriptudgave

Kollationering af filer
Revideret vers. 3

Kollationering af filer foregår i fire trin som beskrevet nedenfor.

1. Præparation af de indgående filer

De indgående filer skal forberedes før tektsammenligningen:

- Visse del af teksten skal ikke kollationeres, fx ikke alternative læsninger eller slettet tekst
- Specielt til *CollateX* skal teksten tokeniseres, dvs. de mindste enheder der skal være genstand for kollationering, skal mærkes
- Tekstens xml-mærkning indgår ikke i kollationen. Hvor dette ønskes, må xml-mærkerne omskrives til tekst.

Det første punkt varetages af et stilark, *filter.xsl*. Heri er det hensigten at projektafhængige filtreringer kan foretages.

Den samlede opgave varetages af en transformation vha. stilarket *prepColx.xsl*., som importerer *filter.xsl*. Dette og de følgende stilark tager som inddata en liste over de filer som skal transformeres, fx har vi i filen *skakt.xml*:

```
<linkGrp type="alignment" n="Oernereden">
  <link xml:id="sktX"
    target="txt007.xml#sktX txt008.xml#sktX txt009.xml#sktX txt010.xml#sktX"
    n="englænderen"/>
  <link xml:id="skt00"
    target="txt007.xml#skt00 txt008.xml#skt00 txt009.xml#skt00 txt010.xml#skt00"
    n="intro"/>
  <link xml:id="skt01"
    target="txt007.xml#skt01 txt008.xml#skt01 txt009.xml#skt01 txt010.xml#skt01"
    n="ved midnat"/>
  <link xml:id="skt02"
    target="txt007.xml#skt02 txt008.xml#skt02 txt009.xml#skt02 txt010.xml#skt02"
    n="efter døden"/>
  <link xml:id="skt03"
    target="txt007.xml#skt03 txt008.xml#skt03 txt009.xml#skt03 txt010.xml#skt03"
    n="nyt fra kattene"/>
</linkGrp>
```

Resultatfilerne anbringes her i en særskilt undermappe *colxsrc* hvorfra de er klar til næste trin. De indgående filer deles i afsnit after skakterne, dvs. der dannes et antal filer svarende til antal filer × antal skakter, her 4 × 5 = 20 filer.

Alle ord mærkes enkeltvis med `<w>`, dvs. der lægges op til en ord-til-ord-kollationering. Input til næste trin er et rent tekstindhold, dvs. xml-mærkerne indgår ikke i kollationeringen. Da vi gerne vil have nogle af disse med, nemlig afsnit og titler, har vi i præparationen mærket dem `<tag>&lt;lb/&gt;</tag>`. Tilsvarende mærkes dokumentets titel `<title>&lt;title&gt; ... &lt;/title&gt;</title>`.

2. Kollation

Til kollationen benyttes programmet *CollateX*. Fra kommandolinjen kaldes:

```
collatex[1] colxsrc/txt007.skt01.xml colxsrc/txt008.skt01.xml colxsrc/txt009.skt01.xml ¬
          colxsrc/txt010.skt01.xml ¬
          -xml -xp "//w|//tag|//title" -f tei -o colxres\skt01.col0.xml
```

De anvendte optioner skal forstås

`-xml`	læs inddatafilerne som xml
`-xp "..."`	dette er en *xpath* for de dele af xml-filen som skal indgå i kollationeringen
`-f tei`	uddata formatteres som en tei-fil
`-o ...`	uddatafil
`-s \skt\nonorm.js`	javaScript der angiver at der *ikke* skal ske nogen normalisering af kildeteksten (i form af store/små bogstaver eller mellemrum)

Flg. kald af programmet vil udskrive et resume af de mulige optioner.

```
collatex -h
```

Der foretages en kollation for hvert skaktafsnit hvorfor resultat bliver et antal filer svarende til antallet af skakter som anbringes i en særskilt undermappe, *colxres*.

For at kunne administrere disse programkald er der konstrueret et stilark, *batch.xsl*, som danner en Dos-batch-fil, `tmp.bat`[2].

3. Efterbehandling af resultatfilen

Den resulterende fil er formateret efter en speciel CollateX norm, ca. således:

```
<cx:apparatus xmlns:cx="http://interedition.eu/collatex/ns/1.0"
              xmlns="http://www.tei-c.org/ns/1.0">
  <app>
    <rdg wit="w1">&lt;title&gt;[forarbejde] Collinske Samling 41,4o-II. Hæfte
      II-4&lt;/title&gt;</rdg>
    <rdg wit="w2">&lt;title&gt;[koncept] Collinske Samling 36,4o- III, Nr.
      82&lt;/title&gt;</rdg>
    <rdg wit="w3">&lt;title&gt;[renskrift] HCA/XVIII-79-A&lt;/title&gt;</rdg>
    <rdg wit="w4">&lt;title&gt;[Førstetrykket, kap. VI-VIII:]
      Ørnereden&lt;/title&gt;</rdg>
  </app>&lt;lb/&gt;&lt;anchor xml:id="skt00" corresp="skakt.xml#skt00"/&gt;
  <app>
    <rdg wit="w2 w1"/>
    <rdg wit="w4 w3">Ørnereden.</rdg>
  </app>
```

Som det første re-aktiveres vinkelparenteserne fra trin 1:

```
<cx:apparatus xmlns:cx="http://interedition.eu/collatex/ns/1.0"
              xmlns="http://www.tei-c.org/ns/1.0">
  <app>
    <rdg wit="w1">
      <title>[forarbejde] Collinske Samling 41,4o-II. Hæfte II-4</title>
    </rdg>
    <rdg wit="w2">
      <title>[koncept] Collinske Samling 36,4o- III, Nr. 82</title>
    </rdg>
    <rdg wit="w3">
      <title>[renskrift] HCA/XVIII-79-A</title>
    </rdg>
    <rdg wit="w4">
      <title>[Førstetrykket, kap. VI-VIII:] Ørnereden</title>
    </rdg>
  </app><lb/><anchor xml:id="skt00" corresp="skakt.xml#skt00"/>
  <app>
    <rdg wit="w2 w1"/>
    <rdg wit="w4 w3">Ørnereden.</rdg>
  </app>
```

Det gøres med stilarket *actang.xsl*, der erstatter alle forekomster af `>` med > og tilsvarende `<`

1 Egl. `\skt\collatex`, på Mac `~/skt/collatex`, se senere under Praxis. Tilsvarende kvalificeres de i det følgende omtalte stilark med mappen `skt`

2 På Mac danner tilsvarende *batch-sh.xsl* en script fil `tmp.sh`

med <.[3]

Derefter sendes alle filerne igennem transformationsstilarket *colx2tei.xsl*, som forsyner dokumentet med en header og ændrer `<cx:apparatus>` til valid tei.

4. Visning i spalter

Dokumentet er nu klar til visning med stilarket *collate.xsl*, som er indsat som stilark for filen; den kan dermed vises direkte i en browser. Desværre tillader de færreste browsere (med *Microsoft Explorer* som en undtagelse, i det mindste op til version 11) at læse xml-filer med tilknyttet xsl-stilark, man må derfor enten anbringe filen på nettet eller man kan transformere filen videre til html med *Oxygen* eller med *xml2html*.

Praxis

Transformationerne med stilarkene kræver en xsl-transformer. Af dem findes adskillige, fx indbygget i *Oxygen*. Alternativt kan hele processen køres fra kommandolinjen.

MS-DOS Command Prompt

Hertil er batchfilen *collate.bat* konstrueret:

```
if not exist colxsrc md colxsrc
if not exist colxres md colxres
call \skt\xslt2 \skt\prepColx.xsl %1
call \skt\xslt2 \skt\batch.xsl %1 > tmp.bat
time /t
call tmp.bat
call \skt\xslt2 \skt\colx2tei.xsl %1
```

Som sagt styres hele forløbet af en skakt-fil, som derfor angives som argument til batch-filen, fx

```
\skt\collate skakt.xml
```

Denne praksis kræver installation af en kommandolinje-xsl-transformer. *xslt2* er en sådan, se nedenfor under Programmer.

Den afsluttende transformation til html kan evt. foretages med batchfilen *xml2html*, fx

```
\skt\xml2html Oernereden.col.xml
```

som danner filen `Oernereden.col.html`

MacOS Terminal (*Unix*)

Hertil er skriptet *collate.sh* konstrueret:

```
#!/bin/sh
if [ ! -d "colxsrc" ] ; then mkdir colxsrc; fi
if [ ! -d "colxres" ] ; then mkdir colxres; fi
~/skt/xslt2.sh ~/skt/prepColx.xsl $1
~/skt/xslt2.sh ~/skt/batch-sh.xsl $1 > tmp.sh
tmp.sh
~/skt/xslt2.sh ~/skt/colx2tei.xsl $1
```

Som sagt styres hele forløbet af en skakt-fil, som derfor angives som argument bagefter, fx

```
~/skt/collate.sh skakt.xml
```

Denne praksis kræver installation af en kommandolinje-xsl-transformer. *xslt2.sh* er en sådan, se

3 Dette er gjort som en del af batchen fra forrige trin

nedenfor under Programmer.

Den afsluttende transformation til html kan evt. foretages med skriptet *xml2html.sh*, fx

```
~/skt/xml2html Oernereden.col.xml
```

som danner filen `Oernereden.col.html`

Programmer

Der er idet foregående omtalt flg. programmer. Alle kan findes sammen med stilark m.v. (dokumentation fx) på *http://etxt.dk/skt/*. Det er i noget af det foregående forudsat at disse filer er installeret i en egen mappe */skt/*, hent den nyeste version (aktuelt *skt.2023.zip*) og pak den ud.

CollateX, et kollationerings- eller sammenligningsprogram fra *The Interedition Development Group*. Se https://collatex.net/, hvor der både er dokumentation og vejledning i download.

Der er tale om et java-program, hvilket vil sige at man skal have installeret en java-processor, alt som forklaret i installationsnoterne. I det ovenstående er forudsat at man formulerer en dos-batch-fil, *collatex.bat*:

```
@echo off
java -jar \skt\collatex-tools-1.7.1.jar %*
```

Den vedlagte xsl-transformator er *Saxon* og er hentet fra https://jar-download.com/?search_box=saxon (23 January 2007). Dokumentation findes på https://www.saxonica.com/documentation9.1/using-xsl/commandline.html. *xslt2.bat* indeholder:

```
@echo off
java -jar \skt\saxon-9.1.0.8.jar %2 %1
```

Saxon kan også anvendes til at transformere xml til html, *xml2html.bat*:

```
java -jar /skt/saxon-9.1.0.8.jar -a -versionmsg:off -s:%1 -o:%~n1.html
```

Særligt for Mac

Det er tilstræbt at de samme procedurer kan gennemføres i en Mac terminal, idet de tilsvarende skriptfiler har postfixet .sh i stedet for .bat og skråstregerne vender den anden vej, fx *xslt2.sh*:

```
#!/bin/sh
java -jar ~/skt/saxon-9.1.0.8.jar $2 $1
```

Det kan være vanskeligt at anbringe sine filer direkte under rodkataloget på en Mac. Til gengæld har man en fiks angivelse af den enkelte brugers område, nemlig en tilde ~, og det foreslås derfor at man anbringer skt-mappen der.

Eftersom skriptsyntaksen er en anelse anderledes hos Mac, er også stilarket der danner batchfilen anderledes, det hedder *batch-sh.xsl*.