RIGSARKIVET

Rigsarkivet har i perioden 2014-2017 deltaget i et EU-projekt kaldet *E-ARK* (*European Archival Records and Knowledge Preservation*) http://www.eark-project.com/

Projektet har blandt meget andet udviklet et *open source* udtræksværktøj, der kan koble sig til flere forskellige *database management* systemer og trække databaser ud til en række forskellige bevaringsformater, herunder udtræk til arkiveringsversioner, der lever op til bekendtgørelse 1007 af 20. august 2010.

Udtræksværktøj - frit tilgængeligt

Dette dokument har det formål at videreformidle *open source* udtræksværktøjet *Database Preservation Toolkit (DBPTK)*. Udtræksværktøjet er et program, der kan koble sig til flere forskellige *database management* systemer - eksempelvis *Oracle* og *Microsoft SQL server* - og trække dataindhold, datastrukturer og metadata fra databaser ud til en række forskellige bevaringsformater - herunder udtræk til arkiveringsversioner, der lever op til strukturbestemmelserne i bekendtgørelse 1007 af 20. august 2010.

Værktøjet er frit tilgængeligt på http://www.database-preservation.com/, hvor man kan læse mere om programmet. På siden er der også et link til *GitHub*, hvor kildekoden ligger frit tilgængeligt: https://github.com/keeps/db-preservation-toolkit. På sidstnævnte link kan man indrapportere fejl og mangler til programmet, selv *uploade* forslag til en forbedret kildekode eller få en mere fyldig beskrivelse og dokumentation af programmets funktioner.

På de følgende sider kan man læse om de mere tekniske detaljer om programmet, så interesserede har mulighed for at vurdere, om værktøjet kan være nyttigt. Vær opmærksom på, at der er tale om et *open source* værktøj, som hele tiden kan være i udvikling, og den nyligst opdaterede dokumentation af funktioner vil pt. være at finde på https://github.com/keeps/db-preservation-toolkit/wiki/Application-usage. Vær også opmærksom på, at selvom værktøjet er funktionelt, så kan det stadig være nødvendigt at foretage yderligere bearbejdning før eller efter udtræk, for at arkiveringsversionen kan godkendes iht. bek. 1007 – dette kan man også læse om på følgende sider.

Rigsarkivet fraskiver sig ethvert ansvar for fejl, som skulle opstå i forbindelse med anvendelsen af programmet. Bemærk, at Rigsarkivet ikke kan besvare spørgsmål og yde support på programmet - spørgsmål til programmet bør og kan stilles på https://github.com/keeps/db-preservation-toolkit.

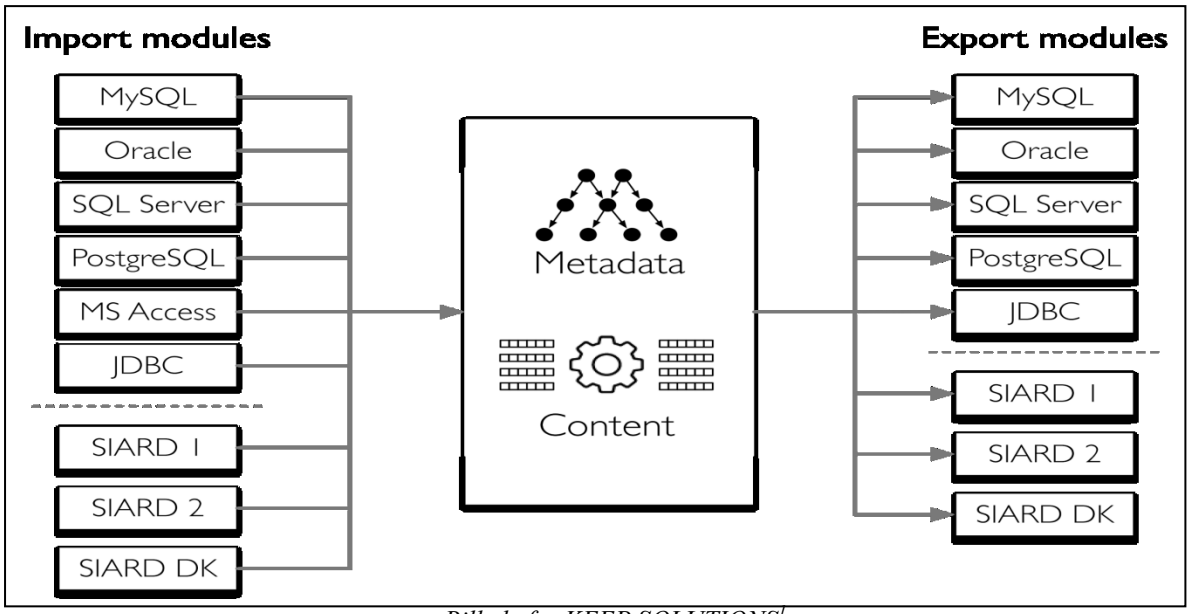
Indhold

1 Generelle tekniske informationer

DBPTK er et java-baseret program, der blot kræver, at Java Runtime Environment er installeret. Værktøjet kan således bruges på styresystemer som Windows, Mac OS, Linux og Solaris. Java Runtime Environment kan hentes på https://www.oracle.com/downloads/index.html.

Der er endnu ikke dannet en brugergrænseflade til programmet, så indtil videre kan man benytte værktøjet via en kommandoprompt.

DBPTK består af to moduler: et importmodul og et eksportmodul. *DBPTK* kan importere og eksportere data, datastrukturer og metadata fra diverse *database management* systemer og bevaringsformater, som kan ses på følgende billede:



Billede fra KEEP SOLUTIONS[1]

Det er således muligt via *DBPTK* at kopiere data fra en kørende database og danne en arkiveringsversion i et af de tre bevaringsformater, som *DBPTK* understøtter.

DBPTK er *open source* og licensrettighederne[2] er GNU Lesser General Public License, version 3.

[1] *KEEP SOLUTIONS (KEEPs)* er et portugisisk firma, der leverer *open source* løsninger til arkivverdenen. *KEEPs* var deltager i E-ARK projektet og har udviklet (og udvikler stadig på) DBPTK http://www.keep.pt/

[2] http://www.gnu.org/licenses/lgpl-3.0.html

2 Bekendtgørelse 1007 = *SIARD DK* (hvad skal man trække ud til?).

SIARD er et akronym for Software Independent Archiving of Relational Databases og "*SIARD 1*" er et åbent schweizisk bevaringsformat, der blev det første bevaringsformat i *SIARD*-"familien". Formatet blev udviklet af *Swiss Federal Archives*[3] og indgik i 2008 som officielt bevaringsformat i det europæiske langtidsbevaringsprojekt *PLANETS*[4].

Rigsarkivet brugte *SIARD 1* som grundstruktur til udarbejdelsen af bekendtgørelse 1007 af 20. august 2010, "Bekendtgørelse om arkiveringsversioner". I internationalt regi er bek. 1007 derfor kendt som *SIARD DK*. Det er derfor SIARD DK modulet man skal bruge for at udtrække til en dansk arkiveringsversion i overensstemmelse med bek. 1007.

I *E-ARK*-projektet har partnerne sammen med *Swiss Federal Archives* udarbejdet *SIARD 2* formatet[5] og samtidig udviklet *DBPTK*, der kan trække data ud til *SIARD 2* formatet. Da *SIARD 1* og *SIARD DK* adskiller sig på nogle få punkter, er det i projektet sikret, at *DBPTK* også kan importere og eksportere fra og til disse.

3 Parametre til modulerne – eksempel på udtræk

For dokumentation af de parametre, som man kan bruge til værktøjet, bør man orientere sig på https://github.com/keeps/db-preservation-toolkit/wiki/Application-usage. Her følger dog et eksempel på et enkelt udtræk fra en testdatabase på en *Microsoft SQL Server* til *SIARDDK*, hvor man har mulighed for at vælge hvilke tabeller, der skal indgå i udtrækket*:*

Via en kommandoprompt afvikles følgende kommando:

```
java "-Dfile.encoding=UTF-8" -jar "C: \Programmes\Database Preservation Toolkit\dbptk-app-
2.0.0-beta7.2.jar" -i microsoft-sql-server -is localhost -idb Northwind -iu RAtest -ip
123456 -ide -e list-tables -ef "H:\Northwindtabeller.txt"
```

Denne kommando kan deles op i følgende:

```
 1              2              3                                4
java "-Dfile.encoding=UTF-8" -jar "C:\Programmes\Database Preservation Toolkit\dbptk-app-
                               5                  6             7           8
2.0.0-beta7.2.jar" -i microsoft-sql-server -is localhost -idb Northwind -iu RAtest
    9       10       11                 12
-ip 123456 -ide -e list-tables -ef "D:\Northwindtabeller.txt"
```

1: Angivelse af, at der afvikles et javaprogram.

2: Angivelse af tegnkodesæt. Det anbefales altid at bruge denne parameter, som angivet i eksemplet.

3: Angivelse af, at der afvikles et javaprogram.

4: Angivelse af destinationen på konkrete javaprogram - sidste nye kan downloades på http://www.database-preservation.com.

5: -i = *importmodul*. Her *Microsoft SQL Server*.

6: -is = *import-server-name*. Her en nem løsning med blot at angive *localhost.*

[3] Se https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html
[4] http://www.planets-project.eu/
[5] http://www.eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20

7: -idb = databasenavn. Her *Northwind*.
8: -iu = *user*. Angivelse af hvilken bruger *DBPTK* skal bruge til at tilgå databasen. *DBPTK* påkræver, at logge på med bruger og kodeord. Her hedder brugeren *RAtest*.
9: ip = *password*. Brugerens kodeord.
10: -ide = *disable encryption*. Det anbefales at bruge denne.
11: -e = eksportmodul. Her angives *list-tables*, hvilket betyder, at der som udtræksprodukt skabes en tekstfil, der indeholder databasens tabeller.
12: -ef = eksportfolder. Her angives destinationen på den tekstfil *DBPTK* danner. Her dannes en tekstfil kaldet "Northwindtabeller" i roden på D-drevet.

Den genererede tekstfil kan redigeres, så man kan fravælge tabeller, der ikke skal indgå i arkiveringsversionen. Efter redigering kan følgende kommando afvikles:

```
java "-Dfile.encoding=UTF-8" -jar "C: \Programmes\Database Preservation Toolkit\dbptk-app-2.0.0-beta7.2.jar" -i microsoft-sql-server -is localhost -idb Northwind -iu RAtest -ip 123456 -ide -e siard-dk -ef "D:\AVID.SA.12223.1" -etf "H:\Northwindtabeller.txt"
```

Denne kommando kan deles op i følgende:

```
 1              2               3                              4
java "-Dfile.encoding=UTF-8" -jar "C:\Programmes\Database Preservation Toolkit\dbptk-app-
                                5                   6              7             8
2.0.0-beta7.2.jar" -i microsoft-sql-server -is localhost -idb Northwind -iu RAtest
    9       10       11                  12                                13
-ip 123456 -ide -e siard-dk -ef "D:\AVID.SA.12223.1"  -etf "D:\Northwindtabeller.txt"
```

1: Angivelse af, at der afvikles et javaprogram.
2: Angivelse af tegnkodesæt. Det anbefales altid at bruge denne.
3: Angivelse af, at der afvikles et javaprogram.
4: Angivelse af destinationen på konkrete javaprogram - sidste nye kan downloades på http://www.database-preservation.com.
5: -i = importmodul. Her *Microsoft SQL Server*.
6: -is = *import-server-name*. Her en nem løsning med blot at angive *localhost.*
7: -idb = databasenavn. Her *Northwind*.
8: -iu = *user*. Angivelse af hvilken bruger *DBPTK* skal bruge til at tilgå databasen. *DBPTK* påkræver, at logge på med bruger og kodeord. Her hedder brugeren *RAtest*.
9: ip = *password*. Brugerens kodeord.
10: -ide = *disable encryption*. Det anbefales at bruge denne.
11: -e = eksportmodul. Her angives "*siard-dk*", hvilket betyder, at der som udtræksprodukt skabes en arkiveringsversion, der lever op til strukturbestemmelserne i bekendtgørelse 1007 af 20. august 2010.
12: -ef = eksportfolder. Her angives i eksemplet at *DBPTK* skal danne en arkiveringsversion med arkiveringsversionsID 12223, og placere denne i roden på D-drevet.
13: -etf = *export-table-filter*. Her angives destinationen på den tekstfil,som *DBPTK* skal bruge som grundlag til at vælge hvilke tabeller, der skal trækkes ud.

4 Hvad kan *DBTPK* ikke?

Det er vigtigt at understrege, at *DBPTK* kun tager de informationer, der er i databasen, som man ønsker at udtrække data fra. Ved aflevering af IT-systemer kan det forekomme, at der ligger informationer uden for selve databasen:

Tabel- og kolonnebeskrivelser

Systemdokumentation. Hvis indholdet af tabeller og kolonner er beskrevet uden for databasen, så vil disse selvfølgelig mangle i den dannede arkiveringsversion. Da tabel- og kolonnebeskrivelser er obligatoriske jf. bekendtgørelsens 6.C, Figur 6.3, vil *DBPTK* når der dannes en tabelindeksfil *tableIndex.xml*, placere en dummytekst i *description*-elementet, der lyder "*Description should be set manually*" for de tabeller og kolonner, hvor metadata ikke eksisterer.

Flere databaser

Et andet eksempel, hvor *DBPTK* ikke kan lave en færdig arkiveringsversion, er, hvis et system består af flere databaser. Her skal der foretages en del bearbejdning enten inden eller efter udtræk. *DBPTK* er ikke udviklet til at importere fra flere databaser, og har som antagelse, at der kun skal trækkes ud fra én database.

Referencer/koder

Hvis der er relationer, som ikke er opmærket i *database management* systemet via *constraints*, eller der er kodede værdier, hvis oversættelse ikke findes i databasen, så vil disse mangle i den dannede arkiveringsversion.

Dokumenter

Hvis en database indeholder dokumenter, indlejrede som *BLOBs*[6], vil *DBPTK* korrekt trække dokumenter ud i den rette mappestruktur og danne et dokumentindeks, *docIndex.xml*. *DBPTK* trækker dog dokumenterne ud som de ligger indlejret med en binær strøm og giver det udtrukne dokument ekstension *.bin*. Da bekendtgørelse 1007 stiller krav til bevaringsformat for dokumenter, er der derfor normalt en filgenkendelse og konverteringsopgave, som skal fuldføres, før at arkiveringsversionen lever op til dokumentkravene angivet i bek. 1007, bestemmelserne i 4.G og 5.E.

Ligger dokumenterne i eksterne siloer med referencer i databasen, så vil der også være et arbejde med at placere dokumenter i den rette mappestruktur, konvertere til bevaringsformat for dokumenter og sørge for at referencerne i udtrækket til tabeldata efterlever bestemmelserne for dokumentreference (dokID).

[6] https://en.wikipedia.org/wiki/Binary_large_object