

Vejledning til deduplikering af dokumenter

Formål

Denne vejledning beskriver Rigsarkivets anbefalede praksis for at håndtere deduplikering af dokumenter i en arkiveringsversion. Deduplikering forstås som fjernelsen af ens kopier af dokumenter i en database, så kun én unik version af dokumentet bevares. Deduplikering kan fx være hensigtsmæssig for at reducere arkiveringsversionens størrelse, hvis det samme dokument er blevet journaliseret på flere sager fjerne redundante data, og dermed effektivisere arkivbrugerens søgning og øvrig behandling af data.

Overordnet fastslås:

- Der er *ingen krav* om deduplikering, men Rigsarkivet opfordrer til det, hvis det modtagende arkiv vurderer, at deduplikering imødekommer hensyn om minimering af lagringsplads og bedre muligheder for effektiv databehandling.
- Deduplikering defineres *ikke* som kassation. Al information bevares i arkiveringsversionen, der er blot tale om at fjerne redundante dokumentkopier af praktiske hensyn.
- To dokumenter opfattes som værende ens, hvis de har den samme MD5-værdi, som beregnes på baggrund af en fils indhold og dens indlejrede metadata. Eksterne metadata (f.eks. filnavn) må godt variere fra fil til fil – så længe MD5-værdien er ens, betragtes filerne som identiske.

Rigsarkivets anviste løsning

Kopier af det samme dokument *kan* undlades fra arkiveringsversionen, så længe det tilbageværende dokument entydigt kan knyttes til alle sammenhænge hvor dokumentet oprindeligt forekom.

Fremgangsmåden er enkelt:

- Angiv entydig docID <did> for hver unikke dokumentfil i docIndex.xml. Kopier af dokumenter udelades fra at blive registreret i docIndex.xml, og fra arkiveringsversionen i det hele taget.
- Dokumentforekomster registreres i dokumentregistreringstabeller jf. bekendtgørelse 128, afsnit 4.A.3. Flere forekomster kan pege på samme dokumentfil i docIndex.xml via nøglen <did>.

Vær opmærksom på, at der skal tages stilling til filnavnet <ofn> af det tilbageværende dokument, hvis flere ens dokumenter har forskellige navne. Der ligger også en individuel vurdering i at veje fordelene ved at deduplikere op mod potentielle ulemper ved ikke længere at kunne sikre parent-opmærkning i docIndex.xml.





Eksempel

En typisk kandidat for deduplikering er sagsbehandlingssystemer, hvor standardbilag (vejledninger, vilkår, samtykkeerklæringer, etc.) vedhæftes mange sager. Det nedenstående, forsimplede scenarie illustrerer hvordan det kan tages højde for i arkiveringsversionen.

To sagsbehandlere LBT og IAM holder hver et møde om open-source office suiten LibreOffice i deres respektive afdelinger. Efterfølgende opretter sagsbehandlerne en sag hver, hvor det samme dokument *LibreOffice, Writer Guide 6.0* vedhæftes. To kopier af dokumentet ligger dermed journaliseret.

På et senere tidspunkt forbereder myndigheden afleveringen til Rigsarkivet. Efter data fra sagsbehandlingssystemet er eksporteret, viser en sammenligning af alle dokumenternes MD5-værdier, at de to dokumentinstanser er ens. For at spare på lagringsplads vælger myndigheden at inkludere én udgave af dokumentet i arkiveringsversionen i stedet for begge udgaver.

Her kobles dokumentet fra tabellen **dokumenter** til tabellen **sager_dokumenter** ved hjælp af nøglen **PK_dokumenter** i kolonnen **dokumenter.id**, og videre til tabellen **sager** ved hjælp af nøglen **PK_sager** i kolonnen **sager.sagsnummer**.

select * from sager_dokumenter			select * from dokumenter				
c1 (FK) sagsnummer INTEGER	c2 (FK) dokid INTEGER	c3 (PK) id INTEGER	c1 dokumentid INTEGER	c2 dato DATE	c3 dokumenttitel CHARACTER VARYING(80)	c4 (FK) lagringsform SMALLINT	c5 (PK) id INTEGER
1014	1	32	1	2007-06-29	GNU GENERAL PUBLIC LICENSE Version 3, 29 June 2007	1	1
1014	2	33	2	2007-06-29	GNU LESSER GENERAL PUBLIC LICENSE	1	2
1015	3	34	3	2009-02-23	Richard Stallman by Anders Brenna 03	1	3
1015	4	35	4	2009-02-23	Richard Stallman by Anders Brenna 01	1	4
1015	5	36	5	2009-02-13	Richard Stallman at Pittsburgh University	1	5
1015	6	37	6	2014-09-12	Richard Stallman - Fête de l'Humanité 2014 - 010	1	6
1016	12	38	11	1991-08-25	Hello everybody out there using minix	1	11
1017	13	39	12	2019-02-02	Overvejelser omkring Åben Kilde	1	12
1018	11	40	13	2019-02-02	aaben_kilde_multimediae.zip	1	13
1019	14	41	14	2018-07-31	LibreOffice, Writer Guide 6.0	1	14
1021	18	42	15	2019-02-02	Debian Popularity Contest	1	15
1018	15	43	99999991	2019-02-02	smime	1	16
1018	17	44	99999991	2019-02-02	Genvej til Debian Popularity Contest	1	17
1021	16	45	99999992	2019-02-02	lukket_kilde_multimediae.zip	1	18
1021	19	46	99999992	2019-02-02	Proprietær Software	1	19
1022	14	47					

select * from sager				
c1 (PK) sagsnummer INTEGER	c2 dato TIMESTAMP	c3 sagstitel CHARACTER VARYING(60)	c4 (FK) sagsbehandlerid SMALLINT	c5 (FK) sagsbehandlerini CHARACTER VARYING(4)
1014	2019-02-02T08:10:11.000Z	Afklaring af licenstyper	152	KT
1015	2019-01-02T08:10:11.000Z	Forberedelse til evt. møde med Richard Stallmann	17	LBT
1016	2018-12-02T08:10:11.000Z	Debat omkring sidste revision af Creative Commons	119	JNW/
1017	2018-11-02T08:10:11.000Z	Åben Kilde bevidsthed i folkeskolen	119	JNW/
1018	2018-10-02T07:10:11.000Z	Første udkast til et Åben Kilde OS	48	IAM
1019	2018-09-09T07:10:11.000Z	Anbefaling af tekstbehandling til den offentlige sektor	17	LBT
1021	2018-07-02T07:10:11.000Z	Udfordringer med lukket kilde programmel og formater	201	GVR
1022	2018-09-09T07:12:11.000Z	Forslag til glosehåndtering i samfundsforvaltningen	48	IAM

Figur 1: skærbillede fra eksempelarkiveringsversion AVID.SA.18008.1