Introduktion
FGS Databas

Introduktion till förvaltningsgemen-samma specifikationer (FGS) FGS Databas

Vägledning och förklaring till de förvaltningsgemensamma specifikationerna

RAFGS6V0.1

UTKAST

Kontakta oss

Information om arbetet med FGS:er
hittar du på vår webbplats:
www.riksarkivet.se/fgs-earkiv

Du kan även kontakta oss via e-post:
ra-fgs@riksarkivet.se

Juni 2019

Riksarkivet

Riksarkivet

UTKAST

Riksarkivet

INNEHÅLLSFÖRTECKNING

1. Inledning ... 1
2. Familjen FGS Databas ... 1
3. Beskrivning av använda standarder ... 1
   3.1 ADDML ... 1
   3.2 SIARD ... 2
4. Överföring av databaser ... 2
   4.1 Omfattning och avgränsningar ... 2
      4.1.1 FGS:en avser relationsdatabaser eller annan tabellstrukturerad information ... 2
      4.1.2 Dokumentation ... 3
      4.1.3 Terminologi för relationsdatabaser ... 3
   4.2 Användning av FGS Databas och användning tillsammans med andra FGS:er ... 3
      4.2.1 Överföring mellan aktiva system och migrering mellan system ... 3
      4.2.2 Överföring till e-arkiv ... 3
5. ADDML-standarden och dess användning ... 3
   5.1 Beskrivning av ADDML-standarden ... 3
      5.1.1 Vad beskrivs i ADDML 8.2. (Huvudstruktur) ... 4
      5.1.2 Tillhörande information – metadata viktig för bevarandet ... 4
      5.1.3 Strukturen för att beskriva flata filer. ... 5
      5.1.4 Generiska element ... 5
   5.2 Data och ADDML ... 6
6. SIARD-standarden och dess användning ... 6
7. Skillnader, för- och nackdelar med ADDML och SIARD ... 6

UTKAST

Riksarkivet

1. Inledning

Denna introduktion är framtagen för att ge en förståelse för vad de förvaltningsgemensamma specifikationerna för databaser *(FGS Databas (Alternativ 1, relationsdatabaser och register)* samt *FGS Databas (Alternativ 2, relationsdatabaser)*) är och vad de kan användas till. Den förklarar också hur de relaterar till andra FGS:er och hur metadata gällande information ur databaser kan registreras på olika ställen.

2. Familjen FGS Databas

Standarder:
ADDML 8.2
SIARD 2.1

Användningsområde:
I familjen FGS Databas ingår två alternativ:

- *FGS Databas (Alternativ 1, relationsdatabaser och register)*, ADDML
- *FGS Databas (Alternativ 2, relationsdatabaser)*, SIARD

I första skedet tas *FGS Databas (Alternativ 1, relationsdatabaser och register)* fram, i ett andra skede kommer *FGS Databas (Alternativ 2, relationsdatabaser)* att tas fram.

I samtliga fall där det finns information lagrad i databasbaserade ursprungssystem är en FGS för databaser en möjlighet för att definiera hur information ur databaser ska paketeras och förses med metadata som möjliggör överföring och e-arkivering.

FGS Databas erbjuder två alternativ som bygger på två olika standarder. Alternativen har tagits fram för att de båda är de-facto standarder. Samtidigt är de olika till sin karaktär och kan vara att föredra utifrån olika situationer och förutsättningar. Uppdelning i två separata specifikationer har gjorts för att underlätta användningen av dem och för att förenkla vid framtida revideringar och uppdateringar av specifikationerna.

FGS:erna för databaser är avsedda att specificera hur informationen i ett databasbaserat system ska kunna exporteras och struktureras vid överföring mellan olika system, exempelvis mellan två olika verksamhetssystem eller mellan ett verksamhetssystem och ett e-arkiv. Man kan använda FGS Paketstruktur tillsammans med ett av FGS Databas-alternativen. Det går även att använda ett av alternativen för FGS Databas separat, t.ex. vid överföring mellan system.

3. Beskrivning av använda standarder

Avsikten är inte att förklara eller utbilda om standarderna. De som vill eller behöver en mer omfattande beskrivning hänvisas till de organ som ger ut och ansvarar för respektive standard.

3.1 ADDML

”Archival Data Description Markup Language” (ADDML) är en standard med ett XML-schema avsett för långtidsbevarande av information ur databaser vilket utvecklats av Riksarkivet i Norge. I dag används ADDML-standarden av riksarkiven i Norge, Sverige och Finland. Standarden ger möjlighet att ange teknisk metadata för de filer som ett databasuttag enligt ADDML består av. ADDML-standarden ger också möjlighet att definiera själva databasstrukturen i form av posttyper (tabeller) och fält (kolumner) samt primärnycklar och främmande nycklar. Förutom detta innehåller ADDML-standarden också generiska element vilket tillåter användare av formatet att definiera egna informationselement och egna värdelistor. En XML-fil i ADDML-format innehåller metadata **om** de

UTKAST

exporterade databastabellerna, men inte själva informationen **ur** databastabellerna. Den information som exporteras ur systemen lagras i form av sekventiella filer med fast fält och postlängd eller som teckenseparerade filer.

3.2 SIARD

SIARD är en standard med ett XML-format som utvecklats av Schweiziska förbundsarkivet för paketering och arkivering av relationsdatabaser. Förutom i Schweiz används standarden bland annat i Danmark, där Rigsarkivet utvecklat en egen variant av SIARD. Standarden har dessutom tillämpats i EU-projekten Planets och E-Ark. SIARD-standarden möjliggör export av databaser till ett generellt format som innehåller nödvändig metadata för att kunna återskapa en relationsdatabas. Formatet definierar tabeller, kolumner, primärnycklar och främmande nycklar. SIARD kan även definiera databasscheman.

Datafilerna lagras som XML-filer. Ett databasuttag enligt SIARD-standarden utgörs av en paketfil i okomprimerat ZIP-format innehållande datafiler i XML-format samt ett antal metadatafiler i XML-format. SIARD-standarden definierar endast metadata relevanta för relationsdatabaser. Det finns få deskriptiva och kontextuella dataelement och det finns inga generiska dataelement som kan användas för att definiera användarspecifik metadata.

4. Överföring av databaser

4.1 Omfattning och avgränsningar

Databaser kräver särskild metadata för att beskriva poster, fält, primärnycklar m.m. utöver de metadata som redan finns i FGS Paketstruktur. I samtliga fall där det hos myndigheter finns information lagrad i databasbaserade ursprungssystem, men då inga tematiska FGS:er finns (som personal, räkenskaper mm), är en FGS för databaser en möjlighet att definiera hur information ur databaser ska paketeras och förses med metadata för att möjliggöra överföring och/eller e-arkivering.

- Denna FGS är en beskrivning av leveranstypen databaser.
- Denna FGS beskriver och specificerar innehåll samt struktur gällande leveranstypen databaser.
- Denna FGS beskriver endast metadata som är specifik för leveranstypen databaser.
- Denna FGS avser inte att definiera metoder för export ur databaser utan definierar metadata som ska åtfölja ett uttag ur databasen i form av strukturerade textfiler.
- Denna FGS behandlar inte den administrativa process som kan föregå överföring från en databas till ett e-arkiv eller hur informationen ska presenteras för återsökning i e-arkiv.

Innan överföring av information från en databas till ett e-arkiv ska ske så kan det krävas ett förarbete i form av ett leveransprojekt och en leveransöverenskommelse.

4.1.1 FGS:en avser relationsdatabaser eller annan tabellstrukturerad information

Huvuddelen av specifikationen avser metadata som behövs för att kunna beskriva filer som exporterats från databaser. Med databaser avses i första hand relationsdatabaser vilka ska överföras och/eller arkiveras som databasuttag.

Beroende på vilken typ av överföring man ska göra kan man välja det alternativ av FGS Databas som passar bäst. FGS Databas är förutom vid arkivering av relationsdatabaser även tillämplig vid arkivering av information av annat slag. Dels kan det gälla information ur äldre typer av databaser vilka i strikt mening inte är relationsdatabaser, men det kan även röra sig om information som

UTKAST

exporterats ur andra system än databashanteringssystem. I princip kan all information som kan struktureras i tabellform representeras och beskrivas enligt denna FGS. Det kan exempelvis vara statistikfiler eller forskningsdata. Det kan också handla om information i Excel-format som behöver exporters till ett icke-proprietärt format för att kunna arkiveras på ett författningsenligt sätt. Om textfiler är ett lämpligt bevarandeformat beror dock på Excelfilernas innehåll. I de fall då Excelfiler exempelvis innehåller formler eller kommentarer så kommer dessa inte att bevaras när filerna omvandlas till sekventiella textfiler.

4.1.2 Dokumentation

I en leverans till e-arkiv enligt denna FGS bör det också ingå grundläggande system-dokumentation, exempelvis i form av diagram som beskriver databasens inbördes relationer och beroenden. Dokumentationen kan lämpligen bifogas som PDF-filer, men det kan även ingå filer i andra filformat. För dessa medföljande dokumentationsfiler sker beskrivningen i FGS Paketstruktur och dess anpassning.

4.1.3 Terminologi för relationsdatabaser

För databaser kan man använda lite olika terminologi beroende på sammanhang. I detta dokument används följande terminologi:

- *Posttyp* (= tabell, relation)
- *Fält* (= kolumn, attribut)
- *Post* (= rad, tupel)
- *Fil* (= uttag från en tabell eller en vy)

4.2 Användning av FGS Databas och användning tillsammans med andra FGS:er

Denna FGS kan för överföring användas separat eller tillsammans med andra FGS:er.

4.2.1 Överföring mellan aktiva system och migrering mellan system

Denna FGS kan användas vid överföring av databasinformation mellan aktiva system. Den kan också användas vid migration av databasinformation, vid exempelvis byte av systemmjukvara. I båda dessa fall går det bra att använda FGS Databas separat. Den metadata som finns på paketnivå kontrolleras i dessa fall i allmänhet på annat sätt och behöver inte förpackas. Om så önskas kan i dessa fall FGS Paketstruktur användas kombinerat med FGS Databas.

4.2.2 Överföring till e-arkiv

Denna FGS kan användas för överföring till e-arkiv, både mellanarkiv och slutarkiv. I dessa fall rekommenderas användning tillsammans med FGS Paketstruktur.

5. ADDML-standarden och dess användning

5.1 Beskrivning av ADDML-standarden

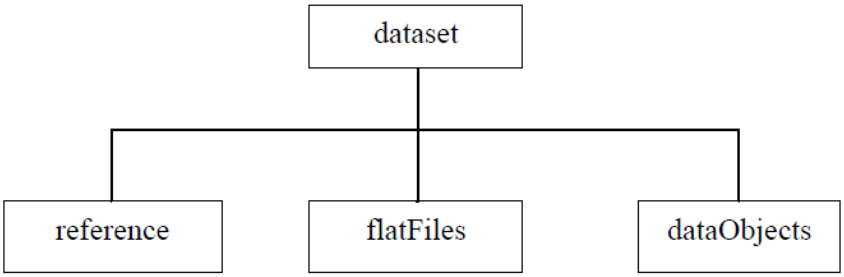
”Archival Data Description Markup Language” (ADDML) utvecklades ursprungligen under 1990-talet av Norska Riksarkivet. Den hette då ADDMML där det extra M:et stod för ”manipulation”. Sedan 2009 utvecklas standarden i avtalat samarbete med svenska Riksarkivet och använts också av

UTKAST

finska Riksarkivet. I *FGS Databas (Alternativ 1, relationsdatabaser och register)* används ADDML 8.2.[1]

5.1.1 Vad beskrivs i ADDML 8.2. (Huvudstruktur)

ADDML består av tre huvuddelar. En del som beskriver tillhörande information, dvs. information om data i databasuttaget (datasetet) - *reference*. En del som är själva beskrivningen av de flata filerna (datafilerna) i detalj - *flatFiles*. Samt en del som ger möjlighet att koppla andra filer än flata filer till beskrivningen - *dataObjects*.

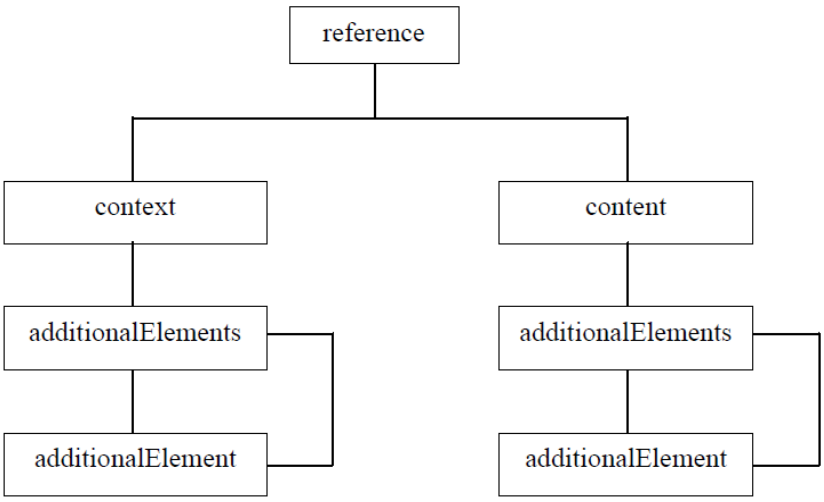


Figur 1. Huvuddelarna i ADDML.

I *reference* kan man registrera metadata som är viktig för bevarandet gällande den kontextuella och innehållsmässiga nivån. I *flatFiles* kan man registrera en detaljerad beskrivning av strukturen i datafilerna oavsett om de är i fast format eller teckenseparerade. Filer som inte är i fast eller teckenseparerat format kan beskrivas i *dataObjects*. Där är det inte möjligt att göra en detaljerad beskrivning.

I *FGS Databas (Alternativ 1, relationsdatabaser och register)* rekommenderas däremot inte användande av *dataObjects*. Motsvarande metadata finns istället i METS-metadata specificerad i FGS Paketstruktur.

5.1.2 Tillhörande information – metadata viktig för bevarandet



[1] ADDML, Archival Data Description Markup Language, Generell del, 2010-09-16

Figur 2. Översikt över elementen i *reference*.

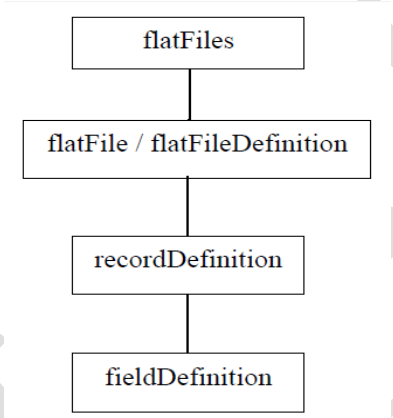
Med tillhörande information menas information om databasuttaget. Detta kan vara upplysningar om vem som har skapat databasuttaget, vilken typ av system det kommer ifrån, bakgrund för data i databasuttaget, osv. Detta är inte beskrivande information av själva databasuttaget.

I ADDML är all sådan information samlad i *reference*. *reference* har två underelement. Dessa beskriver:

- kontextuell information
- innehållsrelaterad information

Grupperingarna har inga fasta element, men man kan definiera egna element som kan eller ska följa med ett databasuttag. Se vidare avsnittet om generiska element.

5.1.3 Strukturen för att beskriva flata filer.



Figur 3. Översikt över element för att beskriva databasuttag (i förenklad form).

Huvudstrukturen för att beskriva en fil har en *flatFiles* som översta nivå. En *flatFiles* kan innehålla en eller flera *flatFile*. På motsvarande sätt kan en *flatFile* innehålla en eller flera *recordDefinition*, och därpå innehåller *recordDefinition* en eller flera *fieldDefinition.*

I en vanlig relationsdatabas kommer det normalt inte behövas postnivå (*recordDefinition*). I andra sammanhang kan emellertid en fil innehålla många olika typer av poster som man kan behöva splittra upp i var sin tabell. Kopplingen mot en databas kommer därför vara mot postnivå och inte mot filnivå. På grund av detta är också all information om nycklar lagd på postnivå.

5.1.4 Generiska element

I ADDML-standarden finns det möjlighet att använda generiska element. Det är element som enbart till namn specificeras i ADDML-schemat och sedan fastslås tekniskt i XML-filen. Ett antal element i *FGS Databas (Alternativ 1, relationsdatabaser och register)* är generiska element och specificeras av FGS-dokumentet.

Det finns i ADDML olika sätt att skapa generiska element. Detta sker dock alltid genom att bygga en slags hierarkisk struktur. I *FGS Databas (Alternativ 1, relationsdatabaser och register)* har det dels använts additionalElements – additionalElement och dels attributet properties – property.

UTKAST

5.2 Data och ADDML

Överföring av datainnehåll från ett databassystem med hjälp av ADDML-formatet bygger på export till filer motsvarande tabeller i databasen. Denna export kan ske till sekventiella textfiler vilka antingen kan vara i ”fixed format” (fasta post och fältlängder) eller i ”delimited format” (poster och fält åtskilda av postavskiljar- och fältavskiljartecken).

För specifika databasuttag kan urval av tabeller ske. Det handlar oftast om att undvika rena systemtabeller.

6. SIARD-standarden och dess användning

Formatet SIARD kommer ursprungligen från Schweiziska Bundesarchiv. Det har varit en del av EU-projekten PLANETS och E-ARK och används av Danska Rigsarkivet. Här finns också en beprövad erfarenhet av formatet och en spridning i Europa. Till SIARD finns utvecklade verktyg på marknaden som delvis automatisk kan exportera och importera databaser hanterade i de fem vanligaste databasprodukterna.

SIARD-formatet bygger på ISO standard SQL och det är en viktig förutsättning för att kunna hantera ett stort antal databashanterare. SIARD-verktygen kan med bakgrund av detta använda standardkopplingar mot databaser vilket ger förutsättningar för transparens i lösningarna. SIARD-formatet gör det möjligt att förutom informationen (innehållet) bevara relationer och viktiga funktioner i relationsdatabaser (logiken med alla relationer och datatyper). I verktygssviterna finns visualiseringsverktyg för enkel sökning. Formatet möjliggör uttag och import till måldatabas för att möjliggöra avancerade sökningar och mer fullständig funktionalitet för arkiverade databaser.

7. Skillnader, för- och nackdelar med ADDML och SIARD

Här sammanfattas och listas några av de skillnader, för- och nackdelar som finns mellan de två standarderna:

- Det finns i Sverige en lång och beprövad erfarenhet av ADDML. Standarden används också i Norge och Finland.
- ADDML tillåter användning av generiska element för att utöka formatet med egendefinierade element och värdelistor.
- Med denna tillämpning av ADDML så möjliggörs en standardisering av begrepp på fält/variabelnivå.
- Med ADDML är det lättare att hantera databaser eller liknande som inte är relationsdatabaser, exempelvis olika typer av register från forskning- och statistikverksamhet.
- Om man vill kunna arbeta oberoende av specifika verktyg, och istället använda de export-möjligheter som normalt finns inbyggda i relationsdatabashanterare, kan ADDML ha fördelar. Det är då lämpligt att separera metadata i XML-format från data i form av sekventiella filer. Dessa sekventiella filer kan relativt enkelt exporteras och importeras av olika relationsdatabashanterare. Det blir också lättare att bearbeta och lämna ut information från enstaka tabeller och slippa hantera hela paket.
- Med ADDML blir det naturligare att vid export välja ut de delar av ett databassystem som man vill bevara. Men med SIARD kan man i princip skapa en rensad tabellstruktur som kopia i ursprungssystemet och sedan exportera den.
- Det är en fördel att det finns öppna export- och importverktyg till SIARD som delvis automatisk kan exportera och importera databaser från/till de fem vanligaste databashanterarna.

UTKAST

- Att använda SIARD resulterar i en total export av ett system. SIARD-formatet gör det möjligt att förutom informationen (innehållet) bevara relationer och viktiga funktioner i relationsdatabaser (logiken med alla relationer och datatyper).
- Till skillnad från ADDML kan SIARD även definiera databasscheman.
- En annan skillnad mot ADDML är att i SIARD lagras även datafilerna som XML-filer.
- SIARD har varit en del av EU-projekten PLANETS och E-ARK och det finns en beprövad erfarenhet av formatet och en spridning i Europa.
- SIARD-formatet bygger på SQL-standarden (ISO/IEC 9075) vilket underlättar för att kunna hantera ett stort antal databashanterare. Verktyg som hanterar SIARD kan med bakgrund av detta använda standardkopplingar mot databaser vilket ger förutsättningar för transparens i lösningarna.
- I verktygssviterna för SIARD finns visualiseringsverktyg för enkel sökning i uttagen.
- SIARD-formatet möjliggör uttag ur ursprungssystem och import till måldatabas vilket ger förutsättningar för avancerade sökningar och mer fullständig funktionalitet för arkiverade databaser.

UTKAST