# Några begrepp

## Pass 1: Inledning och bakgrund

*BAS Online 2021-01-20*

Den här presentationen tar upp några begrepp som du kommer stöta på i BAS Online och förklaringar till hur vi använder dem. Först kommer jag gå igenom vad som menas med forskningsdata och tillgängliggörande. Därefter följer en genomgång av skillnaderna mellan lagring, arkivering och bevarande och till sist kommer vi titta på de viktiga FAIR-principerna.

### Forskningsdata

Vad är det vi menar när vi pratar om forskningsdata? Man kan säga att forskningsdata är det material som har samlats in eller genererats av forskare genom ett projekt för att tjäna som underlag till vetenskapliga analyser och validering av forskningsresultat. Forskningsdata kan vara allt från mätresultat och observationer till datakod, bilder och ljudfiler. De kan bestå av både analog och digital information, men de data som deponeras hos SND är alltid digitala.

### Göra data tillgängliga

Tillgänglighet handlar om hur lätt det är att nå data. När man talar om tillgängliggörande av forskningsdata så innebär det att de data som har samlats in i ett forskningssammanhang placeras i ett sammanhang så att de är synliga och användbara för andra forskare. Data kan göras tillgängliga via ett datarepositorium som SND, via en ämnesspecifik databas, det egna lärosätet eller genom forskargruppen själv. Fast tillgängliggörande innebär inte per automatik att data laddas upp så de kan nås av vem som helst. Forskningsdata kan innehålla känsliga uppgifter, så därför måste man först ta reda på om det finns några juridiska restriktioner, eller etiska begränsningar, som förhindrar att data tillgängliggörs öppet. För en studie med känsliga data så räcker det kanske med att

berätta att studien har gjorts och att data finns. Vill en sekundärforskare sen få tillgång till själva forskningsmaterialet så måste forskningshuvudmannen kontaktas. Huvudmannen brukar vanligtvis vara det lärosäte där forskningen bedrivs. Genom att förse forskningsdata med relevanta metadata och dokumentation kan dataset göras sökbara, vilket ökar möjligheterna för att data ska kunna återanvändas i nya forskningssammanhang och citeras.

Vi kommer senare i BAS Online gå in mer på hur man med hjälp av metadata och metadatastandarder gör forskningsdata mer tillgängliga.

## Lagring, långtidsbevarande och arkivering

Lagring, långtidsbevarande och arkivering används ibland lite slentrianmässigt och blandas gärna ihop. Det finns några avgörande skillnader och jag tänkte att vi ska titta lite på dem. Att förlora sitt datamaterial är något som ingen vill råka ut för, så därför är det viktigt att ha en säker lagring där det görs regelbundna backuper. Likaså bör man försäkra sig om att den lagringsmetod man väljer förhindrar intrång från obehöriga. Man vill ju inte att ett dataset med känsliga personuppgifter ska hamna i fel händer. För att inte riskera att filer försvinner om hårddisken kraschar eller datorn blir stulen så rekommenderas att filer aldrig sparas direkt på den egna datorn. Inte heller bör forskningsdata lagras på USB-minnen, externa hårddiskar eller liknande då det finns risk att lagringsmediet går sönder. Data bör lagras så att de är säkra mot intrång och på en yta där det görs regelbundna backuper. Vi kommer titta mer i detalj på det här i *Pass 2: Säkerhet och personuppgifter.*

Långtidsbevarande av digitala data innebär att man sparar ner filer i filformat som med stor sannolikhet går att använda även i framtiden och som inte kräver någon särskild datorutrustning eller programvara. Till skillnad från pappershandlingar som kan arkiveras vid ett tillfälle och som mår bäst av att ligga stilla och hanteras så lite som möjligt så behöver digitala handlingar konverteras regelbundet. Det vill säga att de sparas

om i nyare format och migreras när ett lagringsmedium riskerar att bli omodernt. Data som inte underhålls kommer förr eller senare bli för gamla för att kunna läsas. Hur vet man då vilka filformat som kommer fungera även i framtiden? Ja, sanningen är att det vet man faktiskt inte. Däremot så finns tre riktlinjer som indikerar om det är större sannolikhet att ett visst filformat också kommer fungera på sikt.

1. Formatet ska vara vanligt förekommande, för ett vanligt format löper nämligen mindre risk att avvecklas.
2. Formatet bör vara leverantörsoberoende för då är man inte beroende av en viss programvara för att kunna öppna och läsa filen.
3. Formatet bör ha en öppen teknisk specifikation, vilket innebär att det inte kontrolleras av en enskild person eller organisation.

I praktiken är det inte alltid möjligt att använda filformat som uppfyller alla dessa kriterier, men försöker man följa riktlinjerna och håller sig någorlunda uppdaterad om internationella råd för långtidsbevarande så är förutsättningarna bättre för att en fil ska kunna läsas i framtiden.

Digitala arkivsystem är arkiv för elektroniskt bevarande, hantering och återsökning av digital information. När det gäller digital arkivering är det viktigt att uppmärksamma att forskningsdata alltid ska arkiveras vid det egna lärosätet i enlighet med Arkivlagen (SFS 1990:782). Det gäller såväl rådatafiler och etiktillstånd som forskningsdokumentation och publicerade resultat. SND tar inte över arkivansvar för något deponerat forskningsmaterial.

## FAIR

FAIR-principerna[1] formulerades i januari 2014 vid en workshop som organiserades av den nederländska ELIXIR-noden och det nederländska eSciencecentret. Trots att FAIR är relativt nytt så har det ändå blivit lite av ett ”buzzword” som dyker upp i alla möjliga sammanhang. Men vad är

FAIR, egentligen? FAIR är en akronym som står för *Findable, Accessible, Interoperable* och *Reusable*. FAIR-principerna innebär att forskningsdata ska vara sökbara, de ska vara tillgängliga och åtkomliga, samt möjliga att åter-använda. För att kunna räknas som FAIR behöver forskningsdata beskrivas av relevanta och rika metadata. I pass 3 kommer du få titta mer ingående på metadata, så för ögonblicket räcker det att känna till att metadata är viktiga för att göra data sökbara och tillgängliga, samt att det finns olika krav för hur metadata ska se ut och användas.

Hur arbetar man för att göra forskningsdata FAIR? Först så ska forskningsdata vara möjliga att hitta, det vill säga: de ska vara sökbara. Ett sätt att göra data sökbara är att placera dem i en för ämnesområdet relevant metadatakatalog och beskriva data med relevanta metadata. Man bör försäkra sig om att format, språk och vokabulärer som används är accepterade och används inom det aktuella ämnesområdet. Här har också forskaren en viktig roll. Eftersom det är forskaren som vet vad forskningsprojektets data innehåller så är det också forskaren själv som bäst kan märka upp data med metadata. Det är då viktigt att tänka på att de metadata som används är så omfattande att andra kan förstå och återanvända datamaterialet. Sedan behöver forskaren se till att data lämnas in till ett datarepositorium som kan märka upp data med en beständig identifierare, eller PID, det vill säga ett permanent och unikt ID-nummer som underlättar för korrekt citering av data.

Nästa del i FAIR är att data ska vara tillgängliga. Precis som jag redan har sagt så är inte tillgängliggörande av data detsamma som att data är fritt och öppet tillgängliga för alla. I vissa fall kan de inte vara det, exempelvis på grund av personlig integritet eller säkerhet. Metadata är däremot alltid publika och fritt tillgängliga. I fall där själva dataresursen inte kan tillgängliggöras öppet så kan man istället använda metadata för att berätta att resursen existerar och presentera klart och tydligt vilka villkor som finns för åtkomst och återanvändning av data.

För att data ska vara åtkomliga behöver både data och metadata vara kompatibla med tillgängliga standarder som till exempel OAI och html. Det här är exempelvis viktigt i de fall då dataset från flera olika forskningsprojekt ska sammanfogas till ett enda dataset. Om inte dataseten följer befintliga standarder kan de inte läsas av datorer och då är det svårt, för att inte säga omöjligt, att kunna sammanfoga dem. Vidare bör standardiserade vokabulärer användas för att dokumentera och kategorisera data, vilket t.ex. underlättar kommunikationen mellan personer som tillhör samma ämnesområde eller när data ska sökas fram.

Slutligen så ska data vara möjliga att återanvända. En förutsättning för det är att data beskrivs med tillräckliga metadata, att metadata kan läsas av både människor och datorer och att det finns tydliga upplysningar om licenser och information om datamaterialets ursprung.

Det här var förklaringar till några av de begrepp du kommer stöta på i BAS Online och också den sista delen av introduktionen i pass 1. I nästa pass kommer du få veta mer om datahantering, datahanteringsplaner och vad man behöver tänka på när personuppgifter förekommer i forskningsdata.

## Referenser

FORCE11. The FAIR Data Principles: https://www.force11.org/group/fairgroup/fairprinciples (Hämtad 2021-01-20)