SND
Svensk nationell datatjänst

Säkerhet och personuppgifter

Pass 2: Datahantering och datahanteringsplaner

BAS Online 2021-01-20

I den här presentationen ska vi ta upp frågor som rör data som behöver särskild hänsyn. Vi har tagit upp dessa tidigare, men vi ska kika på det i lite mer detalj. Vi ska börja med att titta lite mer på datasäkerhet och vad informationsklassning innebär. Sedan ska vi gå vidare med personuppgifter och frågor kring pseudonymisering och anonymisering.

Datasäkerhet innebär helt enkelt att man hanterar data säkert under tiden som projektet pågår. Det är förstås viktigt att datamaterialet finns tillgängligt för dem som faktiskt ska arbeta med det, samtidigt som obehöriga inte ska ha tillgång till det.

Man behöver tänka på säkerhet i tre olika sorters miljöer: de tekniska miljöerna ska vara konstruerade så att bara rätt personer kommer åt materialet. Det handlar t.ex. om hårddiskar och backupsystem. De fysiska miljöerna ska förhindra att någon obehörig kan få tillgång till materialet genom att rum är låsta eller materialet förvaras i kassaskåp. I de administrativa miljöerna ska åtkomst kontrolleras genom kontroll över vem som har nyckel, lösenord till systemen och rättigheter att läsa och skriva i systemen.

Informationsklassning

Nu kommer vi gå in på det här med informationsklassning, eller klassning av informationstillgångar som är det begrepp som Myndigheten för samhällsskydd och beredskap (MSB) använder, och det är ett sätt att säkerställa datasäkerheten. De allra flesta lärosäten har riktlinjer för informationsklassning som bottnar i riktlinjer från MSB.

På MSB:s webbplats *Informationssäkerhet.se* står det ”genom att klassa information kan ni identifiera känslig och kritisk information, och därefter vidta åtgärder så att denna kan få tillräckligt skydd. Ni kan även undvika att information får onödigt överskydd med höga kostnader som följd.” Det handlar alltså om att säkerställa att all information som genereras inom en organisation håller rätt skyddsnivå så att känslig information inte riskerar att försvinna eller hamna i fel händer. Samtidigt så vill man undvika att välja en överdrivet säker lagring eftersom det då blir väldigt dyrt. Som forskare och anställd behöver man känna till sitt lärosätes, eller organisations, riktlinjer så att man kan informationsklassa projektets forskningsmaterial. Då vet man sedan vilka säkerhetsåtgärder man behöver vidta för att uppfylla kraven på säkerhet kring materialet.

När man klassar information så används kriterierna *konfidentialitet*, *riktighet* och *tillgänglighet*. Ibland förekommer också andra begrepp som t.ex. spårbarhet, men jag avgränsar mig här till de som är vanligast förekommande. Med konfidentialitet menas att endast behöriga personer får ta del av informationen. Riktighet innebär att man kan vara säker på att informationen inte förändras av någon obehörig, av misstag eller på grund av systemstörning. Med tillgänglighet menas att information ska kunna utnyttjas så som man behöver, när man behöver det.

Matrisen[1] på nästa sida visar den modell som MSB har tagit fram som hjälpmedel för att bedöma vilken säkerhetsnivå som behövs för olika delar av en organisations samlade information. Pausa gärna presentationen och titta lite närmare på modellen.

		Konfidentialitet	Riktighet	Tillgänglighet
3	Allvarlig Hög skyddsnivå	**K3** Information där förlust av konfidentialitet innebär **allvarlig** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**R3** Information där förlust av riktighet innebär allvarlig **negativ** påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**T3** Information där förlust av tillgänglighet innebär **allvarlig** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.
2	Betydande Utökad skyddsnivå	**K2** Information där förlust av konfidentialitet innebär **betydande** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**R2** Information där förlust av riktighet innebär **betydande** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**T2** Information där förlust av tillgänglighet innebär **betydande** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.
1	Måttlig Grundläggande skyddsnivå	**K1** Information där förlust av konfidentialitet innebär **måttlig** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**R1** Information där förlust av riktighet innebär **måttlig** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**T1** Information där förlust av tillgänglighet innebär **måttlig** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.
0	Ingen Ingen skyddsnivå	**K0** Information där förlust av konfidentialitet **inte medför någon** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**R0** Information där förlust av riktighet **inte medför någon** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.	**T0** Information där förlust av tillgänglighet **inte medför någon** negativ påverkan på egen eller annan organisation och dess tillgångar, eller på enskild individ.

All information ska värderas utifrån varje säkerhetsaspekt. Till exempel kan ett dataset i en studie innehålla information som är mycket kritisk när det gäller riktighet och tillgänglighet, men mindre känslig när det gäller konfidentialitet. När nivåerna är satta blir det lättare att identifiera vilka konsekvenser ett otillräckligt skydd av ett dataset får. Möjliga konsekvenser kan exempelvis gälla ekonomisk förlust, negativ påverkan på den operativa verksamheten, överträdelse av rättsliga krav, att varumärket får minskat förtroende, att det uppstår skada på en annan organisation eller det omgivande samhället eller tillochmed personskada.

En organisationsanpassad klassningsmodell beskriver vad det innebär att konsekvenser bedöms som *måttliga*, *betydande* och *allvarliga* i just din organisation. Skillnaden mellan olika konsekvensnivåer är att om nivån bedöms som måttlig för ett dataset och det exempelvis hamnar i klass 1 så kan data få lagras på arbetsstationens lokala hårddisk eller på flyttbart medium utan restriktioner. Informationen får överföras elektroniskt utan kryptering och får sändas via fax och med post såväl extern som internt. Om materialet istället hamnar i klass 3 så krävs det kanske istället lagring på en fristående server i ett isolerat nät och servern ska i sin tur vara placerad separat och inlåst i ett godkänt serverrum. Beroende på vad man förväntar sig att

projektets data kommer att informationsklassas som så finns det alltså all anledning att tänka igenom det här tidigt i projektplaneringen. Ta reda på vad som gäller inom din organisation så att data får det skydd som krävs.

Personuppgifter

En typ av data som man definitivt behöver ta särskild hänsyn till redan i planeringsstadiet är data som innehåller personuppgifter av något slag. En personuppgift är information som på något sätt kan kopplas till en levande fysisk person. Det kan till exempel handla om namn, personnummer, en bild, en ljud-inspelning eller motsvarande. Sådana data kan bara användas om man har en rättslig grund för behandlingen, vilket i de allra flesta fall inom forskning är allmänt intresse.

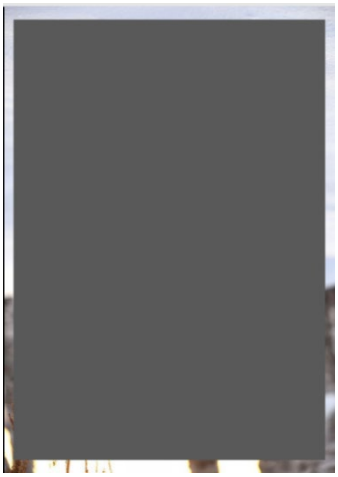
För att dölja personernas identitet kan man koda eller pseudonymisera materialet. Man ersätter helt enkelt direkta identifierare med koder. En kodnyckel finns och förvaras exempelvis inlåst och åtkomlig enbart för en forskare. För de andra i projektet är materialet anonymt: de vet inte vilka personer det handlar om. Men så länge kodnyckeln finns kvar, även om den är svåråtkomlig, så räknas det ändå som personuppgifter: materialet är inte *anonymiserat*, det är bara *pseudonymiserat* eller kodat.

Alltså, pseudonymiserat material är kodat så att det vid direkt anblick inte går att veta vem eller vilka personer som data baseras på. Däremot är det fortfarande personuppgifter eftersom det går att koppla ihop data med en särskild individ genom att använda kodnyckeln.

Anonymiserade data är däremot *inte* personuppgifter. Anonym-isering, eller *avidentifiering*, innebär att man tar bort all möjlig koppling till personerna i materialet. Det går inte med hjälp av några uppgifter som finns i datamaterialet och inte heller med

hjälp av en kodnyckel att koppla ihop uppgifterna i datamaterialet med en person.

För en forskare som t.ex. ska tillgängliggöra sitt material och behöver ta bort personuppgifter handlar det alltså både om att göra sig av med eventuell kodnyckel och att ta bort s.k. identifierare från datamaterialet. Man brukar prata om direkta och indirekta identifierare och jag ska försöka att förklara det här lite mer genom att använda ett ganska simpelt exempel.



Bilden här intill är helt anonymiserad, det vill säga det är i stort sett omöjligt att lista ut vem som döljer sig bakom den gråa rutan. Vi har heller ingen kod som jag kan koppla till en person.

Men om man bara maskerar som på nästa bild blir det lättare gissa vem personen är, trots att personens ansikte är dolt. Du känner till exempel igen skägget och den röda luvan. Att det ligger snö på marken och att ett paket skymtas i säcken hjälper också till. Sådana här uppgifter, som indirekt pekar på en person, brukar man kalla för just indirekta personuppgifter.

Om vi tänker att det här inte hade varit en *bild* på Jultomten utan istället ett frågeformulär som han och hundra andra personer hade svarat på, så skulle det kunna vara så att det finns uppgifter med som gör att man kan avgöra vem som är han. Det

kanske finns frågor om ålder, yrke, bostadsort, osv. i enkäten. För surveydata finns det olika sätt att göra det mindre känsligt och försöka eliminera de indirekta identifierarna. Man kan t.ex. koda om data genom att dela in ålder, längd eller vikt i intervaller så att man tar bort extremvärden. Man kan även koda om uppgifter om hälsotillstånd eller utbildningsnivå. Då blir det svårare att identifiera enskilda personer bland alla respondenter.

Men finns det tillräckligt med uppgifter så är det generellt svårt att säkerställa att materialet är helt och hållet anonymiserat. Då handlar det i slutändan om att avgöra hur *troligt* det är att en individ skulle bli identifierad, och det får man avgöra från fall till fall.

Ett intressant exempel är ett material om ett afrikanskt språk som innehåller en ljudinspelning med en röst som berättar saker. En röst skulle ju i sig kunna vara en personuppgift, om man känner personen som berättar kan man naturligtvis höra precis vem det är. Personen har också på något sätt samtyckt till att inspelningen används. Det har dessutom gjorts bedömningen att sannolikheten att någon skulle identifiera den här personen från den här lilla orten någonstans i Afrika är väldigt liten. Därför bedömdes inte personuppgifter vara något problem i det här forskningsprojektet.

En annan aspekt som kan vara värd att ta upp i det här sammanhanget är att de som deltar i forskning om hotade språk ofta ser sig mer som delaktiga i projektet snarare än som forskningsobjekt. Då vill de finnas med i datamaterialet med namn och bild och få erkännande för att de bidragit till forskningen. Det går emellertid rakt emot tanken om anonymisering eftersom de är tydligt identifierbara och mer eller mindre kräver eller önskar att få synas. Det har därför vuxit fram en tradition inom det här fältet att man ska ge erkännande till informanterna. Det blir med andra ord så att det i vissa

situationer är så offentligt vem som har deltagit att man kanske inte ska diskutera det i termer av nödvändigt anonymiserande.


Jultomten

Åter till vårt exempel med Jultomten. Här har vi både en bild och ett namn, så det råder ingen tvekan om vem personen på bilden är. Det viktiga från det här exemplet är att komma ihåg att det finns en skillnad mellan pseudonymiserat material och anonymiserat material, och att det finns både direkta personuppgifter som brukar vara lätta att hitta och indirekta personuppgifter som inte alltid är lika självklara. Är materialet anonymiserat så är det anonymiserat, för alltid. Men om det är pseudonymiserat så går det att återupprätta personuppgifterna.

Sammanfattning

Den här presentationen har handlat om data som av olika skäl kräver särskild hänsyn med hänseende på datasäkerhet och personuppgifter. Datasäkerhet innebär att man hanterar data säkert under tiden som projektet pågår. Man behöver tänka på säkerhet i tre olika sorters miljöer: tekniska, fysiska och administrativa miljöer. Det handlar om att datamaterialet ska finnas tillgängligt för dem som arbetar med det samtidigt som obehöriga inte ska ha tillgång. För att veta vilken säkerhetsnivå ett material behöver kan man använda sig av informations-klassning.

Om ens datamaterial innehåller personuppgifter finns det flera anledningar till att vara uppmärksam. Direkta personuppgifter pekar direkt på en individ, t.ex. ett namn eller ett personnummer. Indirekta personuppgifter är sådana uppgifter som tillsammans

med annan information gör att man kan lista ut vem det handlar om. Du minns vårt exempel med Jultomten.

Jag tog också upp vikten av att veta skillnaden mellan ett anonymiserat material och ett pseudonymiserat. Är data pseudonymiserade finns det en nyckel sparad någonstans. För anonymiserat material går det inte att härleda tillbaka till undersökningspersonerna.

Referenser

[1]MSB:s matris för klassning av information https://www.informationssakerhet.se/metodstodet/utforma/#klassningsmodell (2020-11-24)