Kravbeskrivning för metadata och data som tillgängliggörs via SND

Innehållsförteckning

Kravbeskrivning för metadata och data som tillgängliggörs via SND

Svensk Nationell Datatjänst (SND) är en nationell forskningsinfrastruktur som från och med 1 januari 2018 drivs av ett konsortium bestående av sju lärosäten[1]. SND har till uppgift att tillgängliggöra forskningsdata och göra dem återanvändbara för forskare[2].

Som medlem i internationella forskningsinfrastrukturer (t.ex. CESSDA ERIC) och certifierat som Trusted Digital Repository enligt CoreTrustSeal-certifieringen[3] behöver SND uppfylla vissa krav på hur forskningsdata och metadata hanteras. Det finns dock inget krav på att de enskilda lärosätena som ingår i SND-samarbetet måste vara certifierade, men kraven som beskrivs här ligger i linje med vissa av kraven för att uppfylla certifiering. Målet på lite längre sikt är att hela SND-samarbetet skall bli certifierat.

Dokumentets syfte

Detta dokument utgör en kravbeskrivning på metadata som ska göras synliga i SND:s metadatakatalog och de forskningsdata som dessa metadata beskriver. Kravbeskrivningen täcker två områden: krav på metadata och kontroll av data. Som komplement till kravbeskrivningen utarbetas en handbok som bland annat kommer att innehålla praktiska beskrivningar för hantering av data och metadata och förslag på arbetsflöden och rutiner som stödjer upprättandet av DAU. Handboken kommer att finnas tillgänglig online.

Krav på metadata

Utifrån de externa krav och referensmodeller som SND har att förhålla sig till så finns det ett antal metadataelement som behövs för att beskriva forskningsdata på ett slags miniminivå. Syftet med denna miniminivå är att forskningsdata som lämnas in till SND ska vara sökbara, tillgängliga och åtkomliga, samt att vidare spridning av metadata ska vara möjlig. Att följa kraven på miniminivå för metadata är ett viktigt steg i arbetet med att uppfylla FAIR-principerna[4].

Miniminivån på metadata säkerställer att de databeskrivningar som publiceras i SND:s nationella metadatakatalog innefattar tillräckliga metadata enligt följande kriterier:

---

[1] Göteborgs universitet, Karolinska institutet, Lunds universitet, Stockholms universitet, Sveriges lantbruksuniversitet, Umeå universitet och Uppsala universitet.

[2] Forskare är SND:s primära målgrupp.

[3] https://www.coretrustseal.org/wp-content/uploads/2018/02/Swedish-National-Data-Service.pdf

[4] FAIR är en akronym som står för Findable, Accessible, Interoperable och Reusable. Se: https://www.force11.org/group/fairgroup/fairprinciples, https://www.dtls.nl/fair-data/fair-principles-explained, och https://content.iospress.com/articles/information-services-and-use/isu824

1. Metadata som utgör en rimlig lägsta nivå för att data ska kunna hittas (lägsta nivå kan skilja sig mellan ämnesområden).
2. Metadata som talar om var och hur data finns tillgängliga.
3. Metadata som är obligatoriska i DataCite (alternativt kan mappas till DataCites obligatoriska krav) och som möjliggör att data kan tilldelas en Digital Object Identifier (DOI).
4. Metadata som är obligatoriska eller centrala i relevanta metadatastandarder (t.ex. INSPIRE, META-SHARE, Data Documentation Initiative) eller hos andra centrala aktörer (t.ex. Clarin, CESSDA).
5. Metadata som är viktiga ur ett administrativt perspektiv för att kunna lokalisera ansvar och möjliggöra förmedling av data.

I kravbeskrivningen listas enbart elementnamn och definition av elementet (svenska och engelska). Övrig information (som till exempel kontrollerade vokabulärer och tillåtna värden, repeterbarhet och villkor) finns specificerad i SND:s metadataprofiler. Miniminivån kommer automatiskt att uppfyllas när de obligatoriska fälten i webbformuläret för att beskriva och lämna in forskningsdata fylls i.

Observera dock att kraven för miniminivå avser inte data som finns beskrivna i andra metadatakataloger, t.ex. TILDA, men vars metadata speglas/skördas till SND:s metadatakatalog. För skördade metadata bör miniminivån ses som ett riktmärke snarare än ett krav och kan behöva justeras från fall till fall.

Miniminivån för metadata är uppdelad i två delar, *gemensamma metadata* och *ämnesanpassade metadata*. Miniminivån för gemensamma metadata gäller för alla typer av data, oavsett ämnesområde, medan ämnesspecifika metadata är anpassade efter ämnesområde. För närvarande finns ämnesanpassade metadata för följande områden:

- Arkeologi
- Medicin och hälsovetenskap
- Miljö-, klimat- och geovetenskaper
- Samhällsvetenskap
- Språkdata

Nedan listas de element som ingår i miniminivån för metadata. Definitioner (svenska och engelska) och eventuella kommentarer finns i tabellen i Bilaga 1.

Gemensamma metadataelement

- Ansvarig institution/enhet
- Beskrivning
- Beständig identifierare
- Distributör

- Huvudman
- Kontaktperson för data
- Nyckelord
- Personuppgifter
- Publiceringsår
- Skapare/primärforskare – person(er) eller organisation(er)
- Språk
- Tillgänglighet
- Titel
- Titel på dataset
- Version

Ämnesanpassade metadataelement

Vissa element förekommer inom mer än ett område.

Arkeologi och historia

- Dataformat/datastruktur
- Geografiskt område
- Tidsperiod(er) för datainsamling
- Tidsperiod(er) som undersökts

Medicin och hälsovetenskap

- Analysenhet
- Dataformat/datastruktur
- Kodnyckel
- Population
- Studiedesign
- Typ av data

Miljö-, klimat- och geovetenskaper

- Dataformat/datastruktur
- Tidsperiod(er) som undersökts
- Minst ett av följande geografiska element
    - Bounding box
    - Geografiskt område
    - Polygon(er)

Samhällsvetenskap

- Analysenhet
- Dataformat/datastruktur
- Geografiskt område
- Insamlingsmetod
- Kodnyckel
- Population
- Tidsdimension
- Tidsperiod(er) för datainsamling

- Tidsperiod(er) som undersökts
- Typ av datakälla
- Urvalsmetod

Språkdata

- Licens
- Typ av data

Kontroll av data

Det är viktigt att forskningsdata som förmedlas genom SND:s metadata-katalog kan förstås och återanvändas av andra. För att kunna garantera detta behövs ett antal åtgärder som redovisas nedan.

Forskningsdata som laddas ned eller förmedlas via SND:s metadatakatalog ska ha genomgått följande kontroller:

- Leveransen är komplett – den innehåller alla data avsedda för förmedling tillsammans med dokumentation som är nödvändig för återanvändning.
- Levererade filer innehåller inga virus.
- Levererade filer går att öppna och läsa.
- Filerna är i ett lämpligt format för återanvändning och tillgängliggörande[5].
- Originalversionen av data finns sparad på en säker lagringsyta.

Notera att kravet är att data måste finnas lagrade på en säker lagringsyta. Däremot föreligger inget krav från SND på någon särskild lagringslösning, eftersom ansvaret för lagring ligger på respektive lärosäte. Observera att för certifiering finns det särskilda krav på lagringslösningar och strukturer vilka SND kan rådge kring om så önskas.

Andra överväganden

Utöver de ovan angivna kontrollerna behöver varje enskilt lärosäte se till att det finns rutiner för att säkerställa att lagring och utlämnande av data uppfyller de krav som ställs i enlighet med Dataskyddsförordningen (GDPR på engelska) och annan relevant lagstiftning. För vissa forskningsdata kan det t.ex. bli aktuellt med sekretessprövning i samband med utlämnande av data.

För återanvändning av forskningsdata är det viktigt att det finns tillräckligt med tillhörande dokumentation. Det är svårt att exakt definiera vad som är tillräcklig dokumentation för återanvändning eftersom det är beroende av bland annat forskningsområde, typ av data och det specifika forskningsprojektet. En vägledning till detta kommer att finnas i DAU-handboken, men det är upp till varje DAU att bedöma om data har den dokumentation som är nödvändig

[5] Se SND:s sida om föredragna filformat: https://snd.gu.se/sv/filformat

för återanvändning. Exempel på dokumentation är variabellista, frågeformulär och teknisk rapport.

Varje lärosäte är själva ansvariga för att utveckla rutiner för hur deponerade forskningsdata lagras och organiseras internt avseende mappstruktur o.dyl. Det är särskilt viktigt att tänka på hur data som ska förmedlas via SND:s tjänster organiseras. I DAU-handboken kommer förslag på lösningar för organisation av data att presenteras.

Bilaga 1 - Miniminivå metadataspecifikation

Tabell 1: Ämnesgemensamma metadataelement

Element sve	Element eng	Definition sve	Definition eng	Kommentar
Ansvarig institution/enhet	Resposible department/unit	Institution/enhet inom huvudmannen, med administrativt ansvar för studiens genomförande.	Department/unit within the organisation with administrative responsibility for the study	Mappas till DataCites obligatoriska element Publisher (om institution saknas mappas istället Huvudman).
Beskrivning	Description	Sammanfattning som beskriver studien och datamaterialet (t.ex. forskningsprojektets huvudsakliga syften, datamaterialets ursprung, egenskap och omfattning, betydande områden som det omfattar etc.)	A summary describing the study and its data (e.g. main purpose of the research project, provenance, nature and scope of the data, major subject areas covered, etc.)	
Beständig identifierare	Persistent identifier	Beständig identifierare (PID), såsom DOI (Digital Object Identifier). Data publicerade via SND får en DOI om de inte redan har en PID.	Persistent identifier (PID), such as DOI (Digital Object Identifier). Data published with SND will receive a PID (DOI) if they do not already have a PID.	Mappas till DataCites obligatoriska element Identifier.
Distributör	Distributor	Organisation med ansvar att skapa/förmedla kopior av data, t.ex. SND, DiVA, Zenodo.	Institution tasked with responsibility to generate/disseminate copies of the data, e.g SND, DiVA, Zenodo.	Genereras av systemet (för data inom SND).
Huvudman	Principal	Organisation med äganderätt och arkivansvar för datamaterialet. Vanligtvis det lärosäte eller annan forskningsorganisation där forskningen/studien genomförts eller där forskaren var anställd.	The organisation that owns and has archival responsibility for the data. Typically, the university or research institute where the study was carried out, or where the researchers were employed.	
Kontaktperson för data	Contact for data	Person/er som kan kontaktas vid frågor om data.	The person(s) who can provide information about the data.	
Nyckelord	Keywords	Nyckelord som hjälper andra att hitta data.	Keywords to help others find the data.	Ämnesspecifika nyckelordslistor kommer att användas.
Personuppgifter	Personal data	Anger om data innefattar personuppgifter. En personuppgift är all slags information som direkt eller indirekt kan hänföras till en fysisk person som är i livet.	Specifies whether data include personal data. Personal data are all kinds of information that is directly or indirectly referable to a natural person who is alive.	
Publiceringsår	Publication year	Det år som data tillgängliggjorts.	The year when the data was made publicly available.	Genereras av systemet (för data inom SND).

				Mappas till DataCites obligatoriska element PublicationYear.
Skapare/Primärforskare -person(er) eller Skapare/Primärforskare - organisation(er)		Den/de personer som är ansvariga för materialet och det intellektuella innehållet av data, och ingår i resursens citering. Organisation eller institution som är ansvarig för materialet och det intellektuella innehållet av data, och ingår i resursens citering. (Elementet används om en organisation, istället för person, ska ingå som skapare i resursens citering.)	Person/people who are responsible for the data material and intellectual content of the data, and are listed in the resource's citation. Organisation or institution responsible for the material and intellectual content of the data, and are listed as creator in the resource citation. (The element is used if an organisation, instead of a person, should be included as creator in the citation).	Mappas till DataCites obligatoriska element Creator.
Språk	Language	Språk för data och tillhörande dokumentation.T.ex. språk på variabelnamn/labels och metadata i datafilen/datafilerna.	Language(s) of the data and documentation. E.g language of the variable names/labels and the metadata within the datafile(s).	CV
Tillgänglighet	Access level	Specificerar tillgänglighetsnivån för datamaterialet.	Specify the level of access to data.	CV
Titel	Title	Titel på studien.	The title of the study.	Mappas till DataCites obligatoriska element Title.
Titel på dataset	Title of the dataset	Ett karakteristiskt, företrädesvis unikt, namn för datasetet. Titeln används i datasetets citering.	A characteristic, preferably unique, name for the dataset. The title is used in the dataset's citation.	Genereras automatiskt till samma värde som elementet Titel. Kan ändras manuellt.
Version	Version	Version av det aktuella datasetet.	The version number of the dataset.	Genereras av systemet (för data inom SND).

Tabell 2: Ämnesspecifika metadataelement

Element sve	Element eng	Definition sve	Definition eng	Ark.	Med.	Milj.	Sam.	Spr.	Kommentar
Analysenhet	Unit of analysis	Beskriver den enhet som analyseras i datamaterialet.	Describes the unit being analyzed in the data collection.	x					CV
Bounding box	Bounding box	Geografiskt område (rektangel) som avgränsas av två longituder och två latituder (öst, syd, väst, nord)	Geographical area (rectangle) defined by two longitudes and two latitudes (east, south, west, north)			x*			*Obligatoriskt att ange något geografiskt element.
Dataformat/ datastruktur	Data format/data structure	Beskriver formatet/strukturen på de data som finns i datasetet, t.ex. numeriska, text, video. Elementet avser inte att beskriva format på datafiler (filformat).	Describes the format/structure of the data included in the dataset, e.g numeric, text, video. The element does not describe the format of datafiles (file format).	x	x	x	x		CV Mappas till DataCites obligatoriska element ResourceType.
Geografiskt område	Geographic area	Geografiskt område som innehållet i data avser.	Geographic area which the data concern.	x		x*	x		CV *Obligatoriskt att ange något geografiskt element
Insamlingsmetod	Mode of collection	Beskrivning av insamlingsmetoden - den process, metod eller teknik som användes för att samla in data.	Description of the method of data collection - the procedure, technique, or mode of inquiry used to attain the data.				x		CV
Kodnyckel	Code key	Data innehåller ID-nummer som kan kopplas till personnummer via en kodnyckel. Kodnyckeln kvarstår även efter forskningsprojektets slut.	Data contains ID numbers that can be linked to social security numbers by using a code key. The code key remains after the end of the research project.		x		x		Relevant för att framöver kunna idenfitiera data vars variabler kan beskrivas i metadataverktyget Register Utiliser Tool (RUT)
Licens	Licence	Om det finns begränsningar för hur data får användas kan de anges med en känd licens (ex. Creative Commons, GPL). Om det som gäller är fri användning sålänge data citeras rekommenderar vi "CC-BY".	Any restrictions on data use may be specified using a well-known license (e.g. Creative Commons, GPL). If you intend free use with a requirement to cite the dataset, we recommend stating ”CC BY”.					x	CV

Polygon(er)	Polygon(s)	Polygon(er) som beskriver den geografiska utbredningen för datasetet.	Polygon(s) that describe the geographic extent of the dataset.			x*			*Obligatoriskt att ange något geografiskt element
Population	Population	Den grupp individer eller objekt som är föremål för forskning och till vilka analytiska resultat refererar.	The group of individuals or other objects that are the object of research and to which any analytic results refer.		x		x		
Studiedesign	Study design	Beskriver forskningsprojektets upplägg och utformning.	Describes the organisation and design of the research project.		x				CV
Tidsdimension	Time Method	Beskriver datainsamlingens tidsdimension.	Describes the time dimension of the data collection.				x		CV
Tidsperiod(er) för datainsamling	Time period(s) for data collection	Tidsperiod när data har samlats in (startdatum och slutdatum, alternativt ”pågående” om datasamlingen inte är avslutad).	Time period during which the data were collected (start and end date, or ongoing if data collection is not completed).	x			x		Datum (ISO 8601)
Tidsperiod(er) som undersökts	Investigated time period(s)	Tidsperiod(er) som innehållet i undersökningen beskriver. Kan avvika från tidsperiod för datainsamling (exempelvis för historiska data).	Time period(s) described. May differ from the time period of data collection (e.g for historical data).	x		x	x		Datum (ISO 8601), alternativt tidsålder (t.ex. Bronsåldern)
Typ av data	Kind of data	Information som beskriver den typ av data som finns i datasetet.	Information describing the kind of data included in the dataset.		x			x	CV Språk: Mappas till DataCites obligatoriska element ResourceType.
Typ av datakälla	Source of the data	Datakälla. Denna kan vara en befolkningsgrupp, ett register, publicerad eller opublicerad datakälla etc.	The source of the data. This may be a population group, a registry, published or unpublished data source, etc.				x		CV
Urvalsmetod	Sampling procedure	Typ av urvalsmetod som användes för datainsamling.	Type of sampling procedure used for data collection.				x		CV