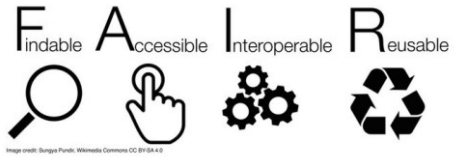


Dokumentation och metadata

Kurs for kursleiarar i FAIR forskningsdatakurering

UiT, Tromsö 20-21 januari 2020

Iris Alfredsson, Svensk Nationell Datatjänst

Del 1:
Generiska metadata

Vad är metadata?

”A statement about a potentially informative object”
Pomerantz, Jeffrey. [2015]. Metadata. Cambridge, MA: MIT Press. S.26

“Structured information that describes, explains, locates, or otherwise represents something else”
National Information Standards Organization [U.S] [2004]
www.niso.org/publications/press/UnderstandingMetadata.pdf

“Metadata is data that is used to describe other data, so the usage turns it into metadata”
Bargmeyer and Gillman [2000]. Metadata Standards and Metadata Registries: An Overview.
https://pdfs.semanticscholar.org/05d6/a22f0ea1da685166787ebed63a44b0ddeac8.pdf

Syftet med metadata

Beskriva data så det går att förstå och använda den både nu och i framtiden

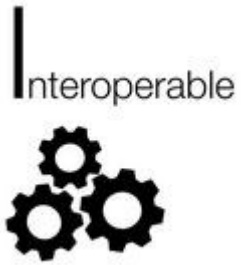
Vad som är metadata ur ett perspektiv kan vara data från ett annat, beroende på användning och på traditioner inom olika ämnesområden

Hur skiljer sig metadata från dokumentation

Metadata är strukturerad information som är lämplig att bearbetas av en dator, men som också kan vara läsbar för människor.

Metadata möjliggör interoperabilitet mellan maskiner, och därigenom kan man skapa ytterligare funktioner med hjälp av informationen.

Bra metadata är standardiserad, konsistent och interoperabel och underlättar sökbarhet, bevarande och arkivering.



I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2 (Meta)data use vocabularies that follow FAIR principles
I3 (Meta)data include qualified references to other (meta)data

Olika typer av metadata

- Beskrivande metadata

Beskriver materialets innehåll och karaktär. Information som kan användas för att hitta data och att bedöma datas kvalitet, t ex titel, skapare, abstract och metod.

- Strukturella metadata

Beskriver strukturen, t ex hur olika delar av ett dataset förhåller sig till varandra.

- Administrativa metadata

Definierar nödvändig information om tekniska egenskaper (tekniska metadata) och rättighetsfrågor (tillgänglighetsmetadata). Denna information är viktig också med tanke på långsiktig bevaring av data (bevarandemetadata).

Beskrivande metadata för citering

Information för att hitta korrekt version av data

- Författare / skapare / primärforskare (kompletterat med ORCID för person)
- Titel
- Publiceringsår
- Version
- Distributör
- Persistent identifierare (PID)

VALU 2019 - SVT:s vallokalsundersökning Europaparlamentsvalet 2019

Citering:

Per Näsman, Sören Holmberg, Henrik Ekengren Oscarsson, Kajsa Gudmundson, Eva Landahl. Sveriges Television AB (2019). *VALU 2019 - SVT:s vallokalsundersökning Europaparlamentsvalet 2019*. Svensk nationell datatjänst. Version 1.0. https://doi.org/10.5878/b6zd-8j63

Metadata för citering i DDI-C

```
<citation>
  - <titlStmt>
        <titl xml:lang="sv">VALU 2019 - SVT:s vallokalsundersökning Europaparlamentsvalet 2019</titl>
        <parTitl xml:lang="en">VALU 2019 - SVT exit poll survey European parliament election 2019</parTitl>
        <IDNo agency="SND">SND 1115</IDNo>
        <IDNo agency="DataCite">https://doi.org/10.5878/pzah-6665</IDNo>
    </titlStmt>
  - <rspStmt>
        <AuthEnty xml:lang="en" affiliation="Royal Institute of Technology, Division of Safety Research">Per Näsman</AuthEnty>
        <AuthEnty xml:lang="sv" affiliation="Kungliga tekniska högskolan, Avdelningen för säkerhetsforskning">Per Näsman</AuthEnty>
        <AuthEnty xml:lang="en" affiliation="University of Gothenburg, Department of Political Science">Sören Holmberg</AuthEnty>
        <AuthEnty xml:lang="sv" affiliation="Göteborgs universitet, Statsvetenskapliga institutionen">Sören Holmberg</AuthEnty>
        <AuthEnty xml:lang="en" affiliation="University of Gothenburg, Department of Political Science">Henrik Ekengren Oscarsson</AuthEnty>
        <AuthEnty xml:lang="sv" affiliation="Göteborgs universitet, Statsvetenskapliga institutionen">Henrik Ekengren Oscarsson</AuthEnty>
        <AuthEnty xml:lang="en" affiliation="Sveriges Television">Kajsa Gudmundson</AuthEnty>
        <AuthEnty xml:lang="sv" affiliation="Sveriges Television AB">Kajsa Gudmundson</AuthEnty>
        <AuthEnty xml:lang="en" affiliation="Sveriges Television">Eva Landahl</AuthEnty>
        <AuthEnty xml:lang="sv" affiliation="Sveriges Television AB">Eva Landahl</AuthEnty>
    </rspStmt>
  - <distStmt>
        <distrbtr xml:lang="en" abbr="SND" URI="https://snd.gu.se">Swedish National Data Service</distrbtr>
        <distrbtr xml:lang="sv" abbr="SND" URI="https://snd.gu.se">Svensk nationell datatjänst</distrbtr>
        <distDate date="2019-08-23">2019-08-23</distDate>
    </distStmt>
    <holdings URI="https://doi.org/10.5878/pzah-6665">Landing page</holdings>
</citation>
```

Andra metadata för sökning

Vad söker vi på?

Exempel på sökning: Jag vill jämföra inställningen till EU-medlemskap i Sverige och Norge på 1990-talet

- Ämne – Titel, abstract, nyckelord
- Geografiskt område – Land, geografisk täckning, Bounding box
- Tidsperiod – Tidsperiod som undersöks, insamlingsdatum
- Datatyp
- …

F1 (Meta)data are assigned a globally unique and eternally persistent identifier
F2 Data are described with rich metadata
F3 (Meta)data are registred or indexed in a searchable resource
F4 Metadata specify the data identifier

Metadata på olika nivåer

Metadata på projekt/studienivå

- Vad är syftet med undersökningen?
- Vad innehåller datasetet?
- Hur samlades data in?
- Vem samlade in data?
- När samlades data in?
- Var samlades data in?
- Hur bearbetades data?
- Etc…

Metadata på datanivå

För numeriska data

Information om datafilen: Format, storlek

Information om variabel: The names, labels and descriptions of variables, their values, a description of derived variables or, if applicable, frequencies, basic contingencies etc. The exact original wording of the question.

Beskrivning av variabel i kodbok

f7: F.7 Förtroende för svenska politiker

Frågetext: Allmänt sett, hur stort förtroende har du för svenska politiker?

Värde	Etikett	Fall	Procentandel
1	Mycket stort	590	6.5%
2	Ganska stort	4845	53.0%
3	Ganska litet	2828	31.0%
4	Mycket litet	870	9.5%
999	Uppgift saknas	123	

Information: Typ: diskret, Format: numeric, Spann: 1-4, Missing: */999

Statistik (Ej vikt./ Vikt.): Valid: (9133 / -) Ej valid: (123 / -)

Beskrivning av en variabel i kodbok skapad i Nesstar

Beskrivning av variabel i DDI xml

```
<var ID="V9" name="f7" files="F1" dcml="0" intrvl="discrete">
    <location StartPos="27" EndPos="29" width="3" RecSegNo="1"/>
    <labl>F.7 Förtroende för svenska politiker</labl>
    <qstn>
     <qstnLit>Allmänt sett, hur stort färtroende har du för svenska politiker?</qstnLit>
    </qstn>
    <valrng>
     <range UNITS="REAL" min="1" max="4"/>
    </valrng>
    <invalrng>
     <item UNITS="REAL" VALUE="999"/>
    </invalrng>
    <sumStat type="vald">9133</sumStat>
    <sumStat type="invd">123</sumStat>
    <catgry>
     <catValu>1</catValu>
     <labl>Mycket stort</labl>
<catStat type="freq">590</catStat>
    </catgry>
    <catgry>
     <catValu>2</catValu>
     <labl>Ganska stort</labl>
     <catStat type="freq">4845</catStat>
    </catgry>
    <catgry>
     <catValu>3</catValu>
     <labl>Ganska litet</labl>
     <catStat type="freq">2828</catStat>
    </catgry>
    <catgry>
     <catValu>4</catValu>
     <labl>Mycket litet</labl>
     <catStat type="freq">870</catStat>
    </catgry>
    <catgry missing="Y">
     <catValu>999</catValu>
     <labl>Uppgift saknas</labl>
     <catStat type="freq">123</catStat>
    </catgry>
    <varFormat type="numeric" schema="other"/>
  </var>
```

Varför är generella metadata viktiga?

“Dublin Core and/or DataCite are especially important because they can provide a bridge between the diverse domains, supporting metadata interoperability as well as data discovery - the F and I of the FAIR Principles.”

SSHOC D3.1 Report on SSHOC (meta)data interoperability problems
https://zenodo.org/record/3569868#.XiQx_EF7k2w

DC – Dublin Core

1995 OCLC/NCSA Metadata Workshop

Dublin, Ohio

fifteen generic elements for describing resources

broad and generic being usable for describing a wide range of resources

https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Qualified Dublin Core included three additional elements (Audience, Provenance and RightsHolder), as well as a group of element refinements (also called qualifiers) that could refine the semantics of the elements in ways that may be useful in resource discovery.

Dublin Core

DC term	Definition
contributor	An entity responsible for making contributions to the resource.
coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
creator	An entity primarily responsible for making the resource.
date	A point or period of time associated with an event in the lifecycle of the resource.
description	An account of the resource.
format	The file format, physical medium, or dimensions of the resource.
identifier	An unambiguous reference to the resource within a given context.
language	A language of the resource.
publisher	An entity responsible for making the resource available.
relation	A related resource.
rights	Information about rights held in and over the resource.
source	A related resource from which the described resource is derived.
subject	The topic of the resource.
title	A name given to the resource.
type	The nature or genre of the resource.

DataCite Metadata Schema

Mandatory properties	
Identifier (with mandatory type sub-property)	M
Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
Titel (with optional type sub-properties)	M
Publisher	M
Publication year	M
ResourceType (with mandatory general type description sub-property)	M

Recommended and optional properties	
Subject (with scheme sub-property)	R
Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
Date (with type sub-property)	R
RelatedIdentifier (with type and relation sub-properties)	R
Description (with type sub-property)	R
GeoLocation (with point, box and polygon sub-properties)	R
Language	O
AlternateIdentifier (with type and relation type sub-properties)	O
Size	O
Format	O
Version	O
Rights	O
FundingReference (with name, identifier, and award related sub-properties	O

DataCite Metadata schema v4.3
https://doi.org/10.14454/7xq3-zf69

Jämförelse DC terms och DataCite terms

DC terms	DataCite terms
contributor	Contributor
coverage	GeoLocation
creator	Creator
date	Date
description	Description
format	Format
identifier	Identifier
language	Language
publisher	Publisher
relation	RelatedIdentifier
rights	Rights
source	-
subject	Subject
titel	Title
type	ResourceType
-	PublicationYear
-	AlternateIdentifier
-	Size
-	Version
	FundingReference

OpenAIRE Guidelines for Data Archives

https://guidelines.openaire.eu/en/latest/data/index.html

- 1. Identifier (M)
- 2. Creator (M)
- 3. Title (M)
- 4. Publisher (M)
- 5. PublicationYear (M)
- 6. Subject (R)
- 7. Contributor (MA/O)
- 8. Date (M)
- 9. Language (R)
- 10. ResourceType (R)
- 11. AlternateIdentifier (O)
- 12. RelatedIdentifier (MA)
- 13. Size (O)
- 14. Format (O)
- 15. Version (O)
- 16. Rights (MA)
- 17. Description (MA)
- 18. GeoLocation (O)

Kontrollerade vokabulärer

Metadata kan kompletteras med en kontrollerad vokabulär (CV - controlled vocabulary).

En kontrollerad vokabulär består av en noggrant utvald uppsättning termer som ska användas för att uttrycka värden i ett metadataelement.

För exempel på kontrollerade vokabulärer som används inom samhällsvetenskap, se CESSDA Vocabulary Service https://vocabularies.cessda.eu/#!discover

ISO-standarder

ISO 8601 är en internationell ISO-standard för angivelse av datum, tid och tidsintervall
20 januari 2020 -> 2020-01-20

ISO 639 är en internationell ISO-standard för språkkoder

ISO 3166 är en internationell ISO-standard för länder

Identifierare för personer och organisationer

ORCID is a nonprofit organization helping create a world in which all who participate in research, scholarship and innovation are uniquely identified and connected to their contributions and affiliations, across disciplines, borders, and time.
https://orcid.org/

ROR is the Research Organization Registry, a community-led project to develop an open, sustainable, usable, and unique identifier for every research organization in the world.
Launched January 2019
https://ror.org/

Del 2:
Ämnesspecifik metadata och dokumentation

Varför är ämnesspecifika metadata viktiga?

"The domain-specific standards provide the refinement that is necessary to satisfy the needs within each domain (the R of the FAIR Principles)."

SSHOC D3.1 Report on SSHOC (meta)data interoperability problems
https://zenodo.org/record/3569868#.XiQx_EF7k2w

R1 (Meta)data have a plurality of accurate and relevant attributes
- R1.1 (Meta)data are released with a clear and accessible data usage license
- R1.2 (Meta)data are associated with their provenance

R2 (Meta)data meet domain-relevant community standards

När, var, hur dokumentera data?

- Beskriv rutiner för dokumentation i datahanteringsplanen
- Tänk igenom vilken information som behövs för att förstå data
- Dokumentera löpande under projektets gång
- Samla informationen på ett strukturerat sätt i en särskild fil, t ex ReadMe-fil, elektronisk loggbok etc.

Exempel på mall för ReadMe-fil se https://data.research.cornell.edu/content/readme

Dokumentation under insamlings- och analysfasen

- Beskriv datainsamlingen: vad, hur, när, vem
- Ange hur data rensats och bearbetats (vad, hur, när, vem)
- Dokumentera vilka datafiler som använts till vilka analyser
- Klargör vilka rutiner som ska följas med hänseende till filnamn, versionering, mappstruktur
- Etc. ….

Resultat från 18 intervjuer inom SSH

"There is a clear division between scientific domains with regard to the most important metadata standards:

- *Social Sciences (CESSDA, ESS, SHARE): DDI Codebook, DDI Lifecycle, DataCite, Dublin Core*
- *Art and humanities (DARIAH): TEI, CIDOC-CRM, Dublin Core*
- *Language science (CLARIN): CMDI, TEI, Dublin Core, OLAC*
- *Heritage science (E-RIHS): EDM, Dublin Core*

The only one mentioned by representatives of all domains is, not very surprisingly, Dublin Core, which is also named as one of the two (in some cases three) most important standards by seven informants."

SSHOC D3.1 Report on SSHOC (meta)data interoperability problems
https://zenodo.org/record/3569868#.XiQx_EF7k2w

Ämnesspecifika metadatastandarder

En metadatastandard består av en uppsättning regler som talar om hur man ska strukturera och formulera metadata.

Standarden innehåller element (bitar av information) som struktureras med hjälp av schema.

RDA Metadata Directory ger en överblick av metadatastandarder för olika områden

https://rd-alliance.github.io/metadata-directory/standards/

Metadataprofil

A profile is typically a subset of a base standard that tailors the metadata elements in the base standard to better describe the data to the community that uses it.

Metadata profiles allow communities to follow a metadata standard, while at the same time enhancing the standard so that it is more appropriate for a particular use or locale.

DDI - Data Documentation Initiative

40-tal medlemsorganisationer (dataarkiv, statistiska centralbyråer, stora forskningsprojekt mm)

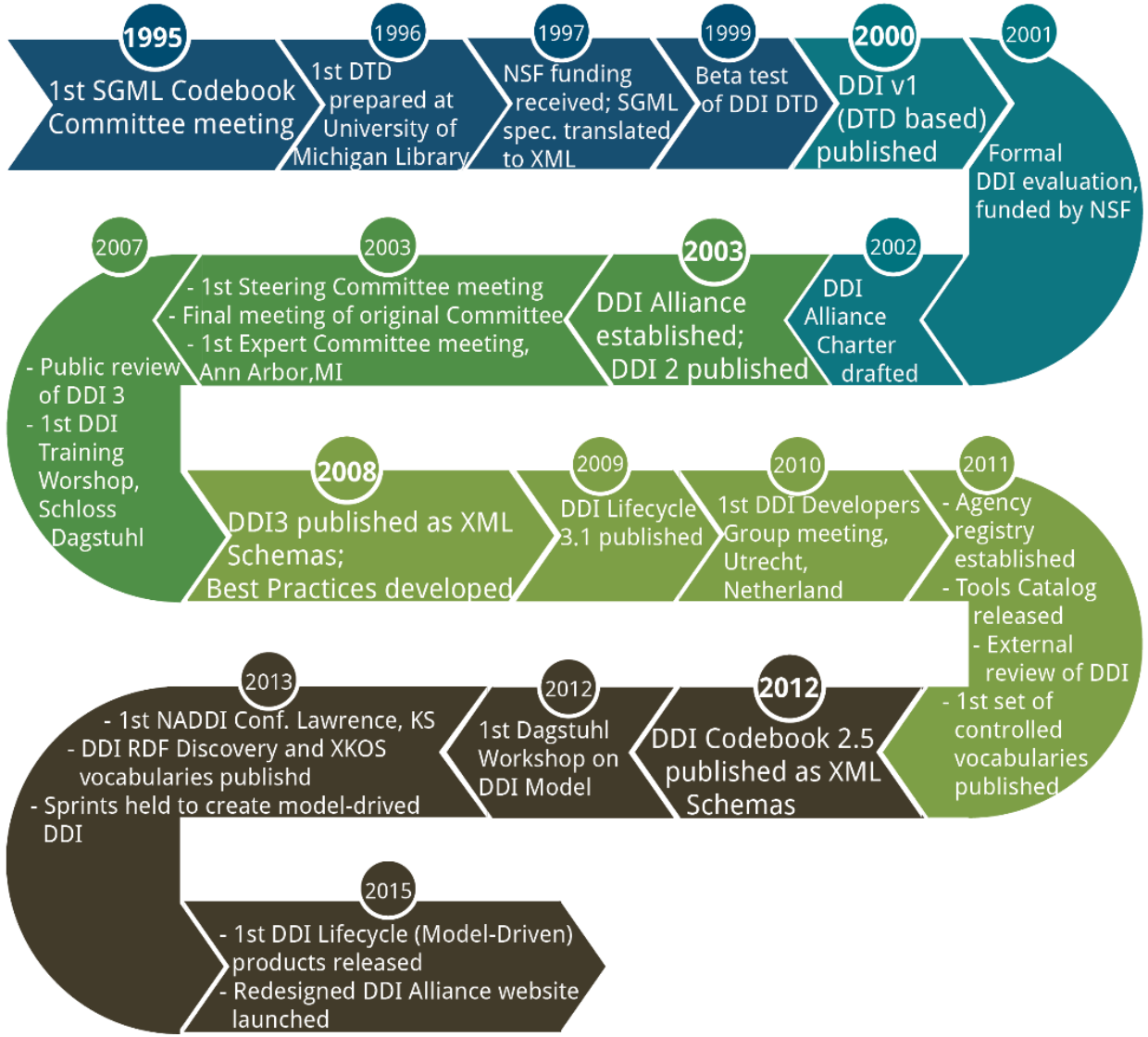
DDI Lifecycle 3.2 (3.3 under 2020)

1154 element

https://ddialliance.org/

Milestones

Data Documentation Initiative



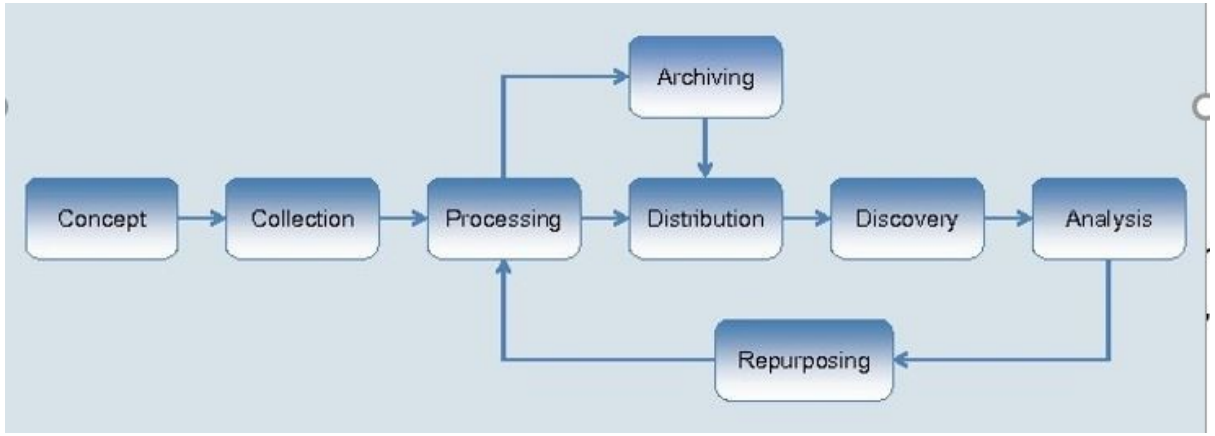
DDI Lifecycle

Första versionen: 2008

Senaste versionen: 2014

Ny version: 3.3 under 2020

1154 element



DDI-Lifecycle is designed to document and manage data across the entire life cycle, from conceptualization to data publication and analysis and beyond.

It encompasses all of the DDI-Codebook specification and extends it.

Based on XML Schemas, DDI-Lifecycle is modular and extensible.

XML schema documentation
https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

DDI Codebook

Första versionen: 2000

Senaste versionen: 2012, modifierad 2014

351 element

DDI-Codebook is a more light-weight version of the standard, intended primarily to document simple survey data. Originally DTD-based, DDI-C is now available as an XML Schema.

XML schema documentation:
https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

Pågående arbete

”DDI Moving Forward” (DDI4)

The DDI metadata standard, originally created to document social science research data, has become relevant to new user groups, including the official statistics and medical research communities.

In order to respond to these new users, DDI is developing a model-based specification (DDI Version 4) that can be expressed in XML Schema, RDF-S/OWL, relational database schema, and program languages.

Such a data model will make it easier to interact with other disciplines and other standards, to understand the specification, to develop and maintain it in a consistent and structured way, and to enable software development that is less dependent on specific DDI versions.

DDI Controlled Vocabularies

Ett 20-tal kontrollerade vokabulärer:

- Analysis Unit
- Data Type
- Mode of Collection
- Sampling Procedure
- Time Method
- Type of Instrument
- …

https://ddialliance.org/controlled-vocabularies

https://www.cessda.eu/

CESSDA Data Catalogue skördar metadata i DDI Lifecycle och DDI Codebook från Service Providers
https://datacatalogue.cessda.eu/

CESSDA Metadata Model (CMM) anger vilka element som är obligatoriska, rekommenderade och frivilliga. Anger också vilka kontrollerade vokabulärer som ska användas.

CESSDA Vocabulary Service är ett verktyg för att hantera och översätta vokabulärer
https://vocabularies.cessda.eu/#!discover

CESSDA EuroQuestion Bank (EQB) kommer 2020. Frågebank som bygger på DDI Lifecycle.

CESSDA Social Science Multilingual Thesaurus är en flerspråkig (fn 14 språk) thesaurus för samhällsvetenskap (ELSST)
https://elsst.ukdataservice.ac.uk/