Implementing coherent metadata for production and dissemination

Introduction

- In recent years, Statistics Denmark has worked on implementing consistent and coherent metadata to support metadata driven production by consolidating Statistics Denmark's statistical metadata into one system.
- The implementation follows international standards and focuses on increasing efficiency, by using standardised metadata actively in the production of official statistics and for dissemination to users.
- The content and organisation of structural metadata aligns with the recommendations and standard outlined in UNECE's Generic Statistical Information Model (GSIM).
- The GSIM model excels by supporting the information objects that go into and out of the individual process steps of the Generic Statistical Business Process Model (GSBPM).
- The paper presents Statistics Denmark's new and improved documentation portal, which showcase relationships between statistical documentation (SIMS quality reports), classifications and code lists, as well as the documentation of data series and variables.

Vision and purpose

- Meet the needs of internal and external users of metadata in relation to their desired use of official statistics (fit-for-purpose)
- Ensure that the quality of statistical products and processes are described and comply with requirements
- Conduct systematic quality assurance of statistical products and processes, and implement quality improvements
- Increase efficiency so that also employees of Statistics Denmark can use shared metadata to complete their tasks, by maintaining metadata in one metadata system and making it easily accessible to all

Two types of metadata

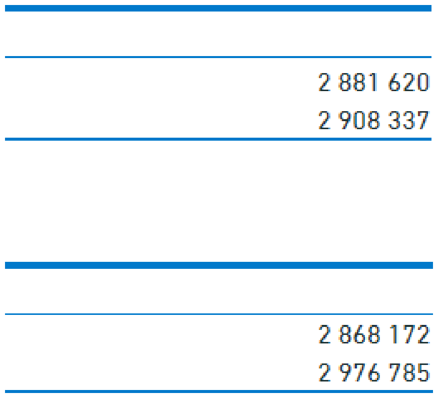
Structural metadata

- Identify statistical data, e.g. titles, code list, time, unit etc.
- Must go together with statistical data
- Impossible to interpret statistics without it

Reference metadata

- Describes statistical concepts and methodologies
- Can be detached from the statistical output
- Statistical documentation (SIMS) is a type of reference metadata

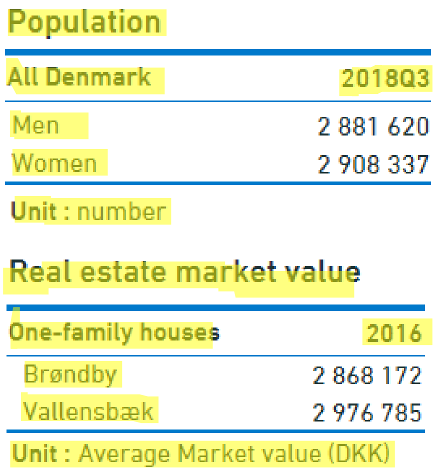
Statistics without metadata



	2 881 620
	2 908 337

	2 868 172
	2 976 785

...with structural metadata



Population

All Denmark	2018Q3
Men	2 881 620
Women	2 908 337

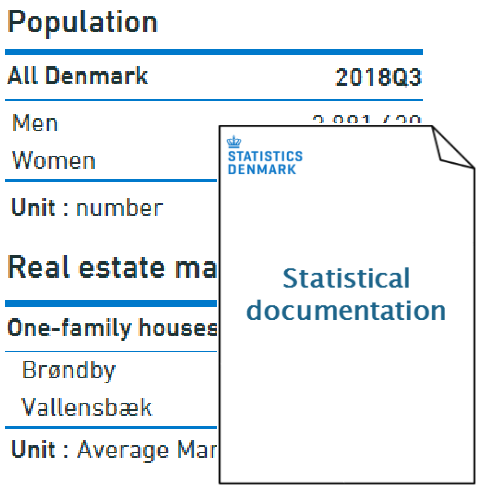
Unit : number

Real estate market value

One-family houses	2016
Brøndby	2 868 172
Vallensbæk	2 976 785

Unit : Average Market value (DKK)

...and reference metadata



Population

All Denmark	2018Q3
Men	2 881 620
Women	

Unit : number

Real estate ma

One-family houses
Brøndby
Vallensbæk

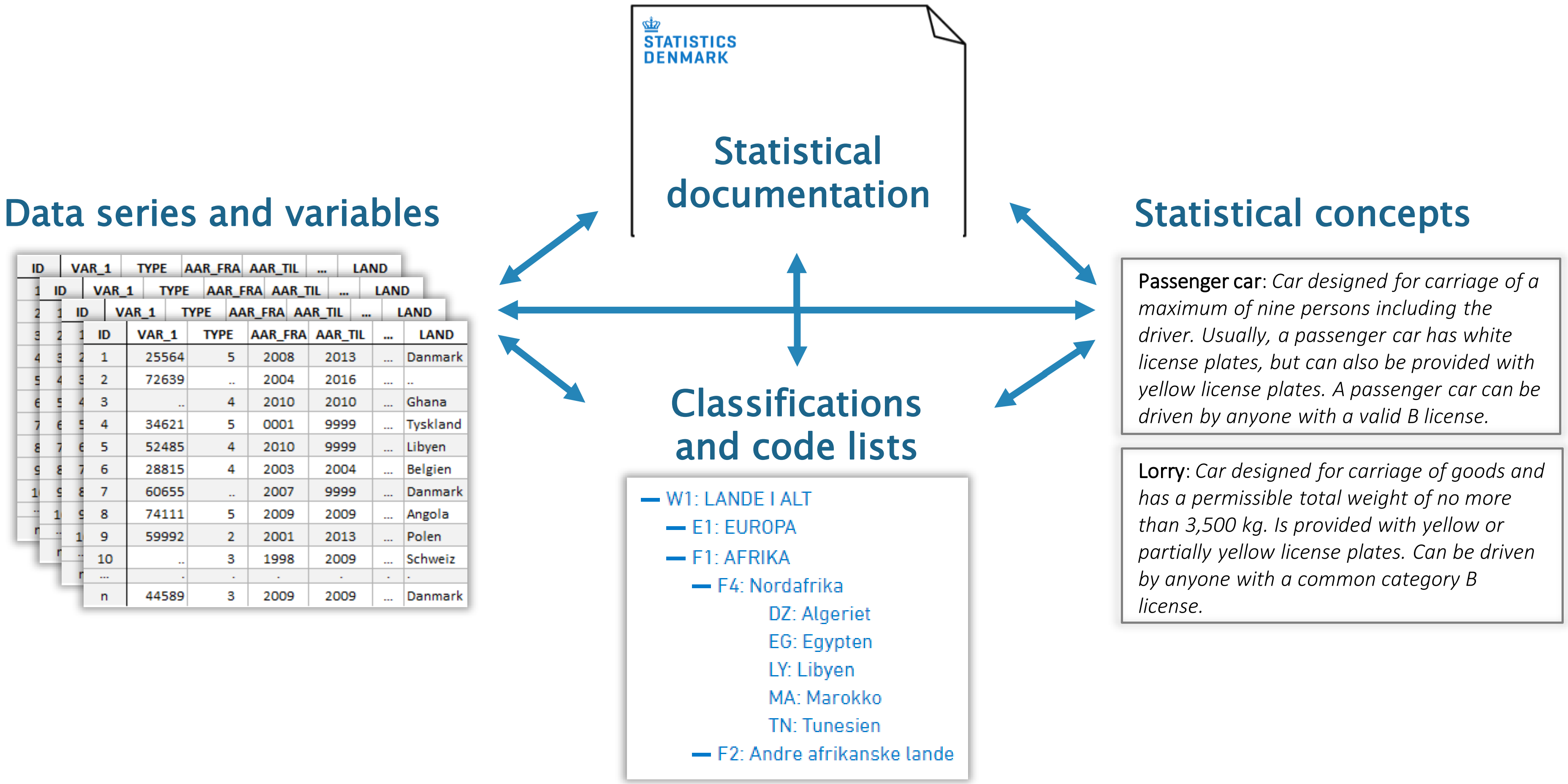
Unit : Average Mar

STATISTICS DENMARK

Statistical documentation

From StatBank FOLK1A og EJDFOE1
Eurostat, Source: https://ec.europa.eu/eurostat/data/metadata/metadata-structure

Four pools of statistical metadata



STATISTICS DENMARK

Statistical documentation

Data series and variables

ID	VAR_1	TYPE	AAR_FRA	AAR_TIL	...	LAND
1	25564	5	2008	2013	...	Danmark
2	72639	..	2004	2016	...	..
3	..	4	2010	2010	...	Ghana
4	34621	5	0001	9999	...	Tyskland
5	52485	4	2010	9999	...	Libyen
6	28815	4	2003	2004	...	Belgien
7	60655	..	2007	9999	...	Danmark
8	74111	5	2009	2009	...	Angola
9	59992	2	2001	2013	...	Polen
10	..	3	1998	2009	...	Schweiz
...	.	.	.	.	.	.
n	44589	3	2009	2009	...	Danmark

Statistical concepts

Passenger car: *Car designed for carriage of a maximum of nine persons including the driver. Usually, a passenger car has white license plates, but can also be provided with yellow license plates. A passenger car can be driven by anyone with a valid B license.*

Lorry: *Car designed for carriage of goods and has a permissible total weight of no more than 3,500 kg. Is provided with yellow or partially yellow license plates. Can be driven by anyone with a common category B license.*

Classifications and code lists

- W1: LANDE I ALT
  - E1: EUROPA
  - F1: AFRIKA
    - F4: Nordafrika
      - DZ: Algeriet
      - EG: Egypten
      - LY: Libyen
      - MA: Marokko
      - TN: Tunesien
    - F2: Andre afrikanske lande

Aligning metadata with international standards

Statistical documentation

- In 2014, statistical documentation (quality reports) were made available online for all statistical products, in both Danish and English and are fully compliant with the Single Integrated Metadata Structure (SIMS).

Classifications and code lists

- Deployed in 2018, classifications and code lists comply with the Neuchâtel terminology model for classification object types and their attributes.

Statistical concepts

- Complying with ISO 704 *Terminology work – principles and methods*, shared quality assured one-place-only and once-only descriptions of statistical concepts, imbedded in statistical documentations, was deployed in 2020.

Data series and variables

- Being deployed at present. We are currently moving existing documentation from our old system to our new system Colectica. The tricky part is aligning objects with the Generic Statistical Information Model (GSIM). The key modernising feature is the introduction of the variable cascade, which splits documentation of variables into conceptual variables, represented variables and instance variables.

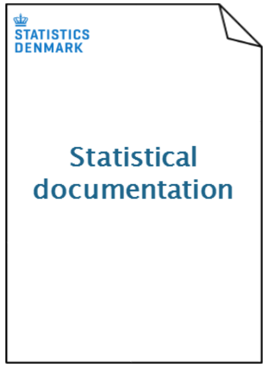
The IT-side of things

- User interface/editor for developers and producers
- Standards based
- Has an API

- Data Documentation Initiative (DDI)
- Relational standard to describe data
- Fokus on reusability

- Relational database
- High performance
- Open source

Quality reports (SIMS)

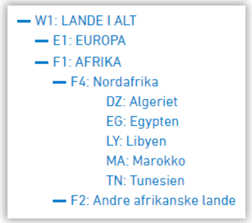

STATISTICS DENMARK

Statistical documentation

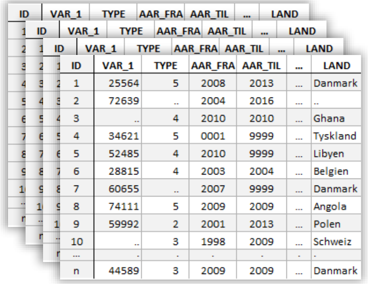
Statistical concepts

Passenger car: *Car designed for carriage of a maximum of nine persons including the driver. Usually, a passenger car has white license plates, but can also be provided with yellow license plates. A passenger car can be driven by anyone with a valid B license.*

Classifications and code lists


- W1: LANDE I ALT
  - E1: EUROPA
  - F1: AFRIKA
    - F4: Nordafrika
      - DZ: Algeriet
      - EG: Egypten
      - LY: Libyen
      - MA: Marokko
      - TN: Tunesien
    - F2: Andre afrikanske lande

Data series and variables



ID	VAR_1	TYPE	AAR_FRA	AAR_TIL	...	LAND
1	25564	5	2008	2013	...	Danmark
2	72639	..	2004	2016	...	..
3	..	4	2010	2010	...	Ghana
4	34621	5	0001	9999	...	Tyskland
5	52485	4	2010	9999	...	Libyen
6	28815	4	2003	2004	...	Belgien
7	60655	..	2007	9999	...	Danmark
8	74111	5	2009	2009	...	Angola
9	59992	2	2001	2013	...	Polen
10	..	3	1998	2009	...	Schweiz
...	.	.	.	.	.	.
n	44589	3	2009	2009	...	Danmark

Quality assurance of metadata

Statistical documentation

- Each publication of statistics must be covered by updated reference metadata, wherein the content, processing and five quality dimensions (relevance, accuracy and reliability, timeliness and punctuality, comparability and accessibility and reliability) is elaborated in detail.

Classifications and code lists

- Ad hoc, but equally systematic in its practice. Whenever a classification or code list needs to be created or updated, a peer who specialises in compliance with Neuchâtel, assists the responsible statistician with understanding, which classification object types are necessary, to load a classification into Colectica.

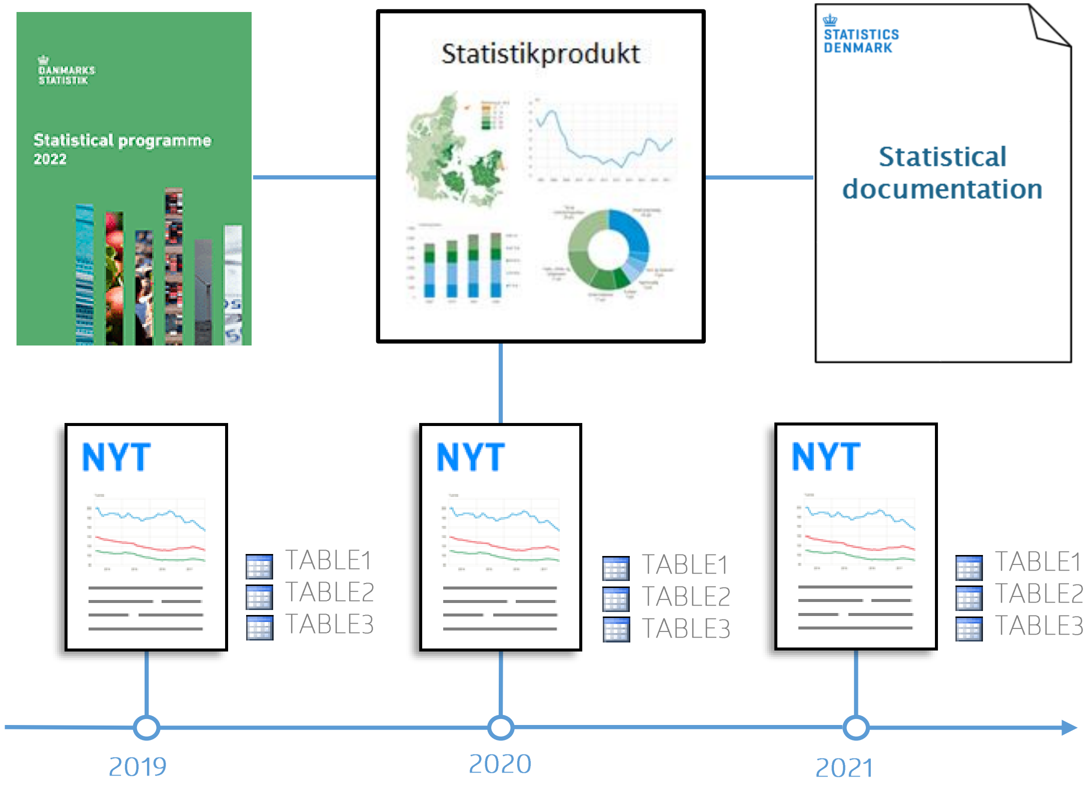
Statistical concepts

- Statistical concepts are meant to be shared between statistical products. Therefore, they must rely on agreed upon definitions and be described in an appropriate manner. We rely on eight “rules” for writing good statistical concept descriptions. 1) Write short and simple, one sentence if possible; 2) consider the target audience; 3) clarity, do not use specialist language, 4) coordinate mutually with other concept descriptions, 5) write adequate, not to narrow or broad, 6) no circularity, 7) no negative definitions and 8) write in singular terms.

Data series and variables

- Within documentation of data series and variables, the establishment of quality assurance procedures is a work-in-progress, to be decided later this year.

Statistical products



OECD definition

"Statistical products are, generally, information dissemination products that are published or otherwise made available for public use that

- describe,
- estimate,
- forecast,
- or analyse

the characteristics of groups, customarily without identifying the persons, organisations, or individual data observations that comprise such groups".

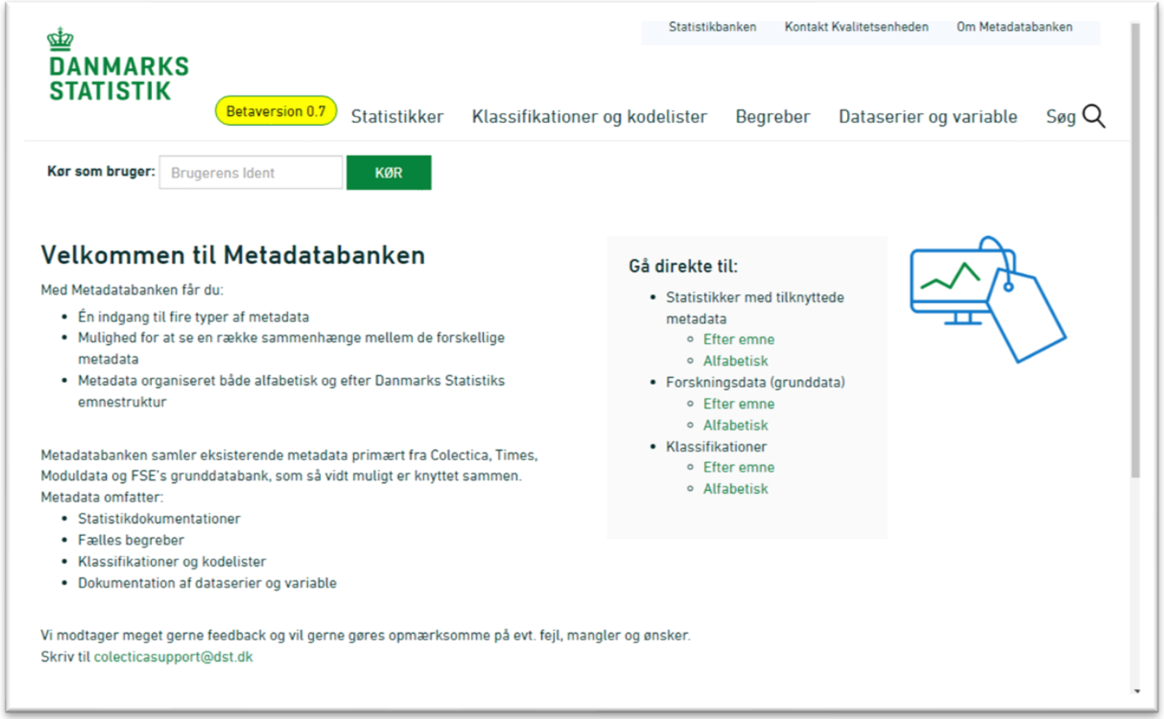
OECD Glossary of Statistical Terms - Statistical Products

Results so far

Launch of the internal documentation portal (1/2)

A clickable webpage that enables user to browse through statistical metadata

Front page



Velkommen til Metadatabanken

Med Metadatabanken får du:

- Én indgang til fire typer af metadata
- Mulighed for at se en række sammenhænge mellem de forskellige metadata
- Metadata organiseret både alfabetisk og efter Danmarks Statistiks emnestruktur

Metadatabanken samler eksisterende metadata primært fra Colectica, Times, Moduldata og FSE's grunddatabank, som så vidt muligt er knyttet sammen.
Metadata omfatter:

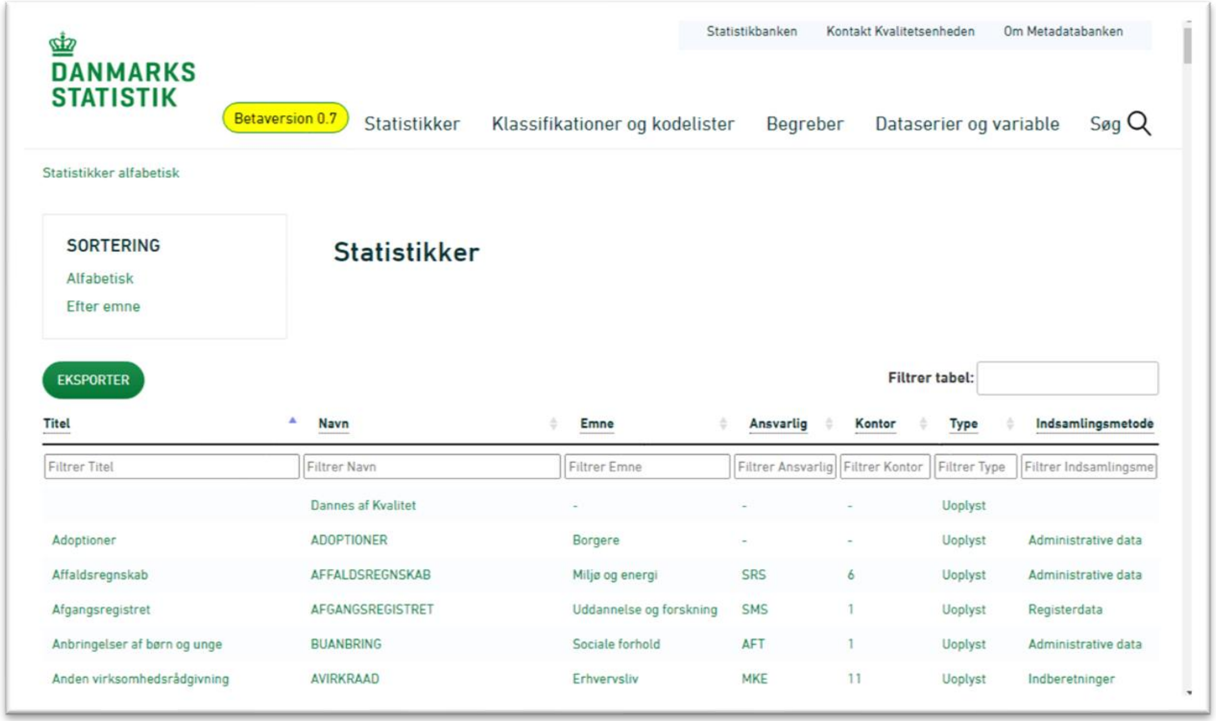
- Statistikdokumentationer
- Fælles begreber
- Klassifikationer og kodelister
- Dokumentation af dataserier og variable

Vi modtager meget gerne feedback og vil gerne gøres opmærksomme på evt. fejl, mangler og ønsker.
Skriv til colecticasupport@dst.dk

Gå direkte til:

- Statistikker med tilknyttede metadata
  - Efter emne
  - Alfabetisk
- Forskningsdata (grunddata)
  - Efter emne
  - Alfabetisk
- Klassifikationer
  - Efter emne
  - Alfabetisk

Browse statistical products



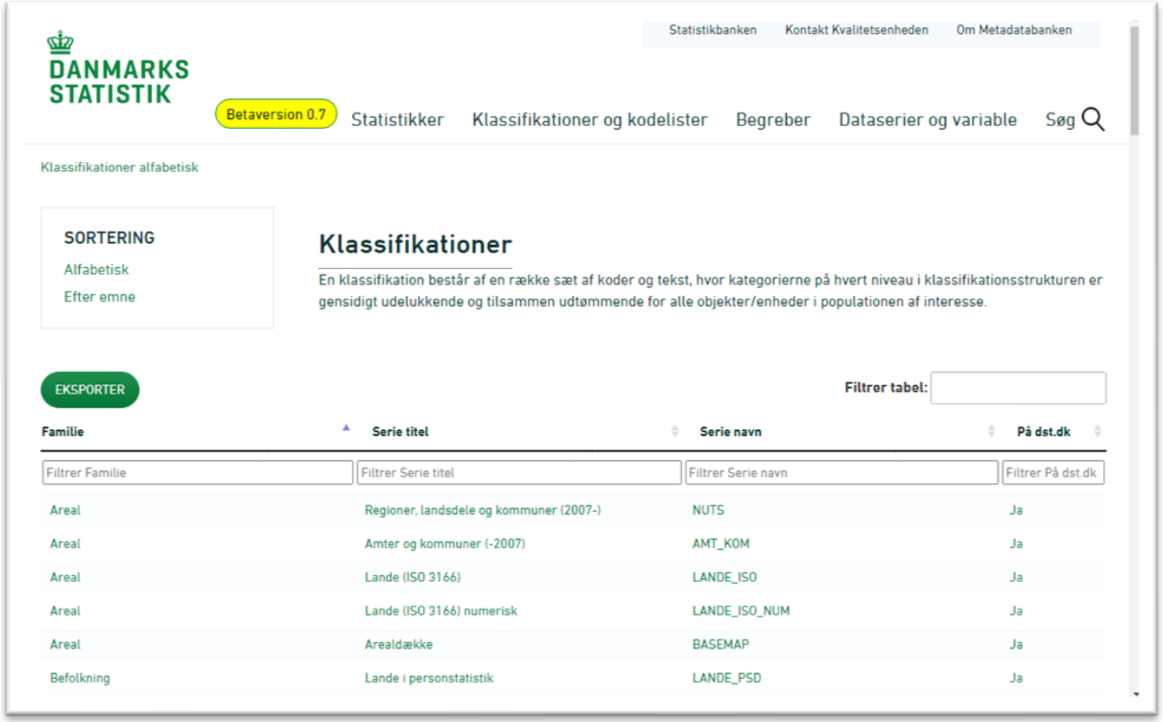
Statistikker

Titel	Navn	Emne	Ansvarlig	Kontor	Type	Indsamlingsmetode
Filtrer Titel	Filtrer Navn	Filtrer Emne	Filtrer Ansvarlig	Filtrer Kontor	Filtrer Type	Filtrer Indsamlingsme
	Dannes af Kvalitet	-	-	-	Uoplyst	
Adoptioner	ADOPTIONER	Borgere	-	-	Uoplyst	Administrative data
Affaldsregnskab	AFFALDSREGNSKAB	Miljø og energi	SRS	6	Uoplyst	Administrative data
Afgangsregistret	AFGANGSREGISTRET	Uddannelse og forskning	SMS	1	Uoplyst	Registerdata
Anbringelser af børn og unge	BUANBRING	Sociale forhold	AFT	1	Uoplyst	Administrative data
Anden virksomhedsrådgivning	AVIRKRAAD	Erhvervsliv	MKE	11	Uoplyst	Indberetninger

Launch of the internal documentation portal (2/2)

A clickable webpage that enables user to browse through statistical metadata

Browse classifications

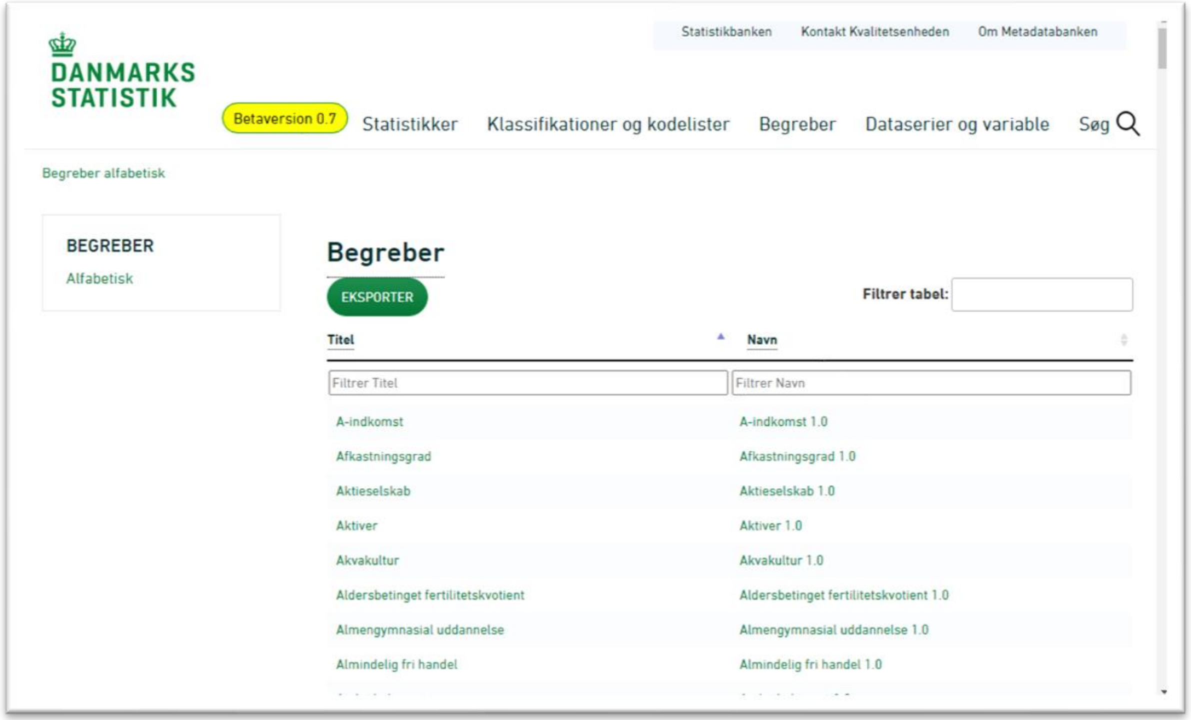


Klassifikationer

En klassifikation består af en række sæt af koder og tekst, hvor kategorierne på hvert niveau i klassifikationsstrukturen er gensidigt udelukkende og tilsammen udtømmende for alle objekter/enheder i populationen af interesse.

Familie	Serie titel	Serie navn	På dst.dk
Filtrer Familie	Filtrer Serie titel	Filtrer Serie navn	Filtrer På dst.dk
Areal	Regioner, landsdele og kommuner (2007-)	NUTS	Ja
Areal	Amter og kommuner (-2007)	AMT_KOM	Ja
Areal	Lande (ISO 3166)	LANDE_ISO	Ja
Areal	Lande (ISO 3166) numerisk	LANDE_ISO_NUM	Ja
Areal	Arealdække	BASEMAP	Ja
Befolkning	Lande i personstatistik	LANDE_PSD	Ja

Browse statistical concepts



Begreber

Titel	Navn
Filtrer Titel	Filtrer Navn
A-indkomst	A-indkomst 1.0
Afkastningsgrad	Afkastningsgrad 1.0
Aktieselskab	Aktieselskab 1.0
Aktiver	Aktiver 1.0
Akvakultur	Akvakultur 1.0
Aldersbetinget fertilitetskvotient	Aldersbetinget fertilitetskvotient 1.0
Almengymnasial uddannelse	Almengymnasial uddannelse 1.0
Almindelig fri handel	Almindelig fri handel 1.0

Metadata linkage from data to output

Micro data

ID	VAR_1	TYPE	AAR_FRA	AAR_TIL	...	LAND
1	25564	5	2008	2013	...	Danmark
2	72639	..	2004	2016	...	..
3	..	4	2010	2010	...	Ghana
4	34621	5	0001	9999	...	Tyskland
5	52485	4	2010	9999	...	Libyen
6	28815	4	2003	2004	...	Belgien
7	60655	..	2007	9999	...	Danmark
8	74111	5	2009	2009	...	Angola
9	59992	2	2001	2013	...	Polen
10	..	3	1998	2009	...	Schweiz
...	.	.	.	.	.	.
n	44589	3	2009	2009	...	Danmark

Variables

Can assume a set of values. Some variables are specified in corresponding code lists, eg TYPE

Variable	Description
VAR_1	...
TYPE	Type of vehicle
AAR_FRA	...
AAR_TIL	...
...	...
LAND	Country

Classification

- W1: LANDE I ALT
  - E1: EUROPA
  - F1: AFRIKA
    - F4: Nordafrika
      - DZ: Algeriet
      - EG: Egypten
      - LY: Libyen
      - MA: Marokko
      - TN: Tunesien
    - F2: Andre afrikanske lande

Codelist

TYPE	TYPE_TXT
1	Passenger car
2	Bus
3	Lorry
4	Truck
5	Motorbike
6	Tractor
7	Autocamper
8	Other

Concepts

Passenger car: *Car designed for carriage of a maximum of nine persons including the driver. Usually, a passenger car has white license plates, but can also be provided with yellow license plates. A passenger car can be driven by anyone with a valid B license.*

Lorry: *Car designed for carriage of goods and has a permissible total weight of no more than 3,500 kg. Is provided with yellow or partially yellow license plates. Can be driven by anyone with a common category B license.*

Statistical Product

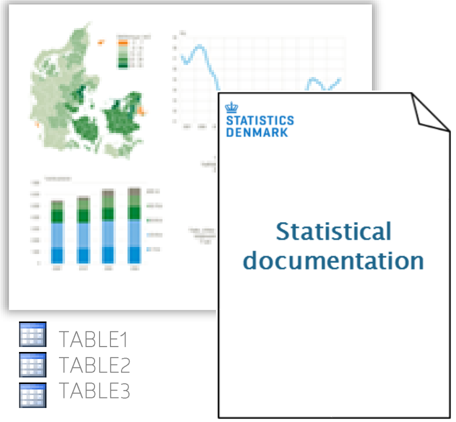


TABLE1
TABLE2
TABLE3

Unit of measure

Antal
Procent
Andel
Kroner
Indeks
Gennemsnit
...

Kroner
1.000 kr.
Mio. kr.
Mia. kr.
...

Looking ahead

Non-exhaustive headlines of work for the next couple of years

- Introduce unit types into our metadata model and link to relevant objects
- Add a user interface layer to the documentation portal in order to edit metadata
- Introduce quality assurance procedures for data series and variables
- Expand the pool of quality assured metadata in Colectica
- Use metadata from Colectica, e.g. directly in questionnaires
- Further integrate metadata from Colectica and the StatBank
- Launch the documentation portal for external users of metadata
- Enhance the use of Colectica API for internal and external system users
- Link GSIM based metadata to GSBPM
- Adopt the FAIR principles for metadata

Thank you for listening!

- Questions?