

Micro Data services to researchers in Statistics Denmark

Activity B.7: Building and integrating data catalogue, data library and documentation

28-30 January 2018

Karin Holst Duer, Statistics Denmark, khd@dst.dk

Charlotte Leolnar Reif, Statistics Denmark, clr@dst.dk



Agenda 28-30 January 2018

Agenda January 28, 2018

- 9:00 – 9:30 Welcoming
- 9:30 – 11:00 **BC**: Expectations, current status, visions
- 11:15 – 12:15 **MS**: Introduction (module data, diamond model status etc.)
- 13:15 – 15:15 **MS**: Presentation of variable catalogue and High Quality Documentation (HQD)

Agenda January 29, 2018

- 9:00 – 9:30 First day review
- 9:30 – 11:00 **MS** and **BC**: Pilot 1 – using the Danish HQD
- 11:15 – 12:30 **MS** and **BC**: Pilot 2 – using the Danish HQD
- 13:15 – 15:00 **MS** and **BC**: Pilot 3 – using the Danish HQD
- 15:15 – 16:00 Ad hoc meetings and additional demonstration as needed

Agenda January 30, 2018

- 9:00 – 9:20 Second day review
- 9:20 – 14:30 **MS** and **BC**: Outlining a data catalogue policy, structure and content of data. Discussing a process for building a catalogue
- 14:30 – 16:00 Summary, evaluation and final remarks

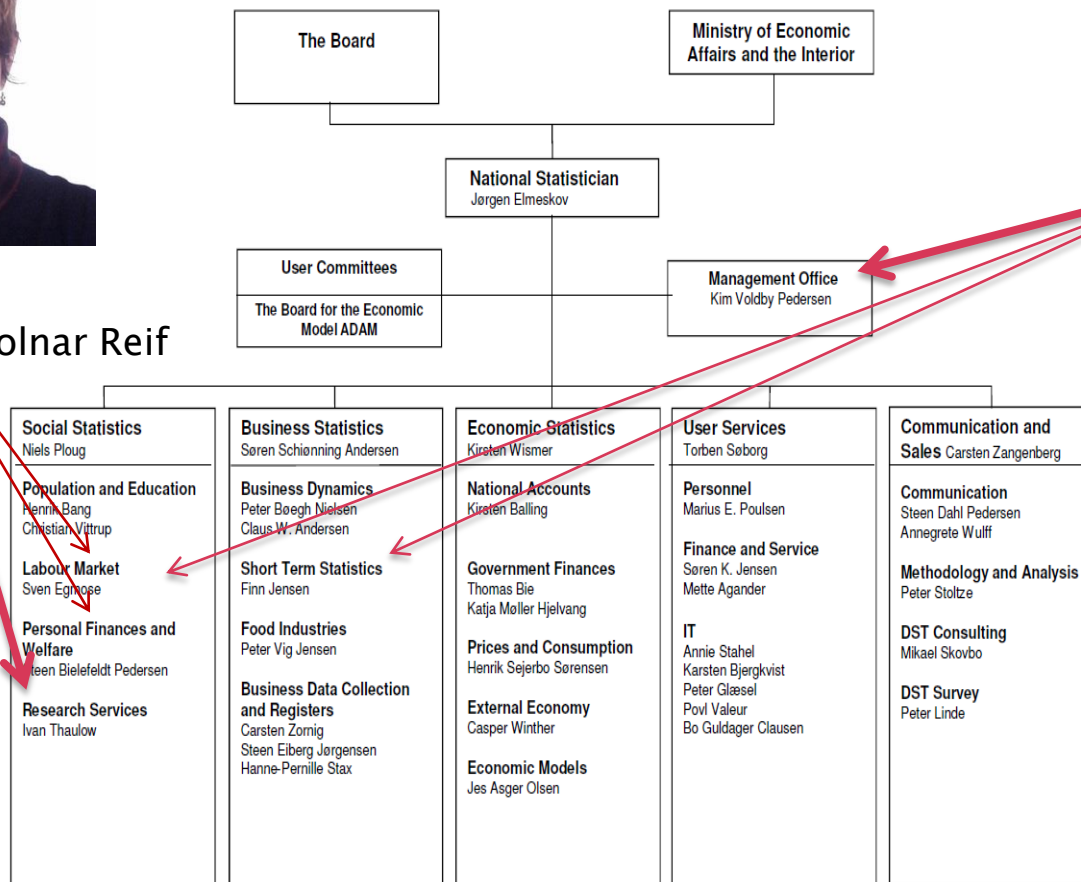
Presentation by MS



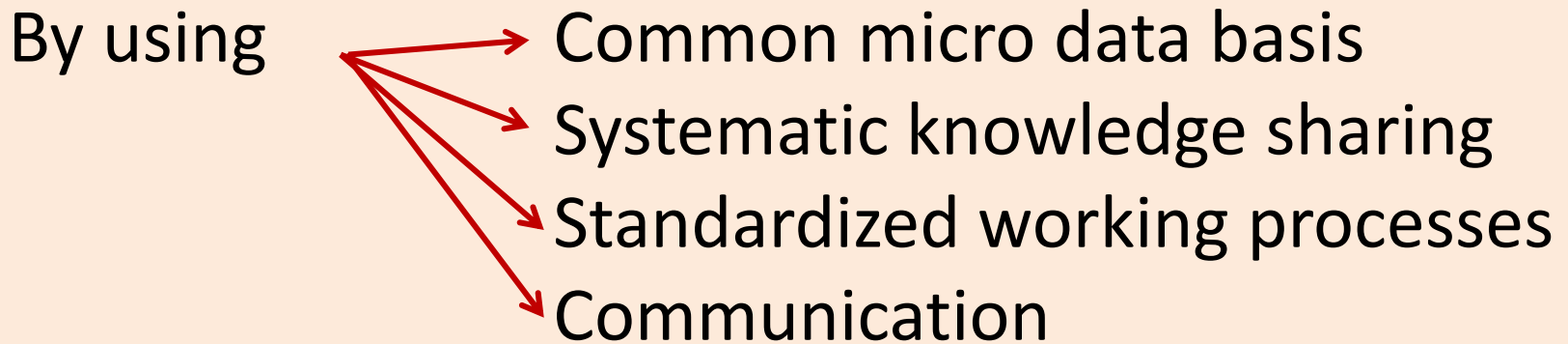
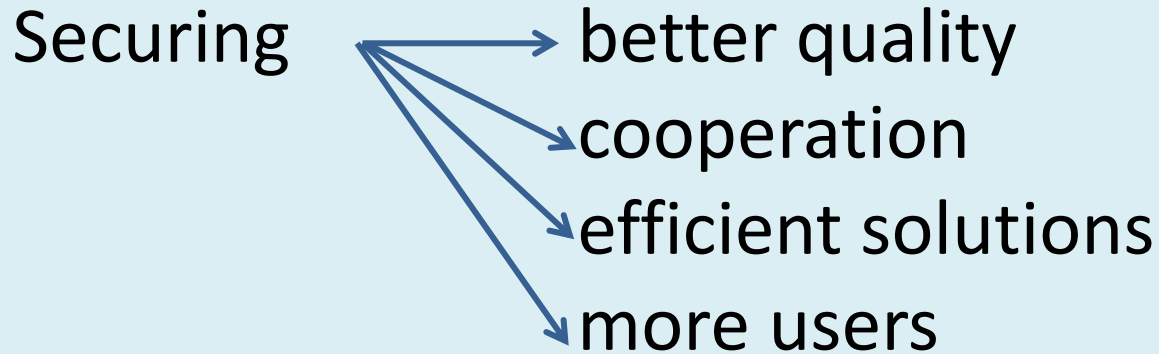
Charlotte Leolnar Reif



Karin Holst Duer



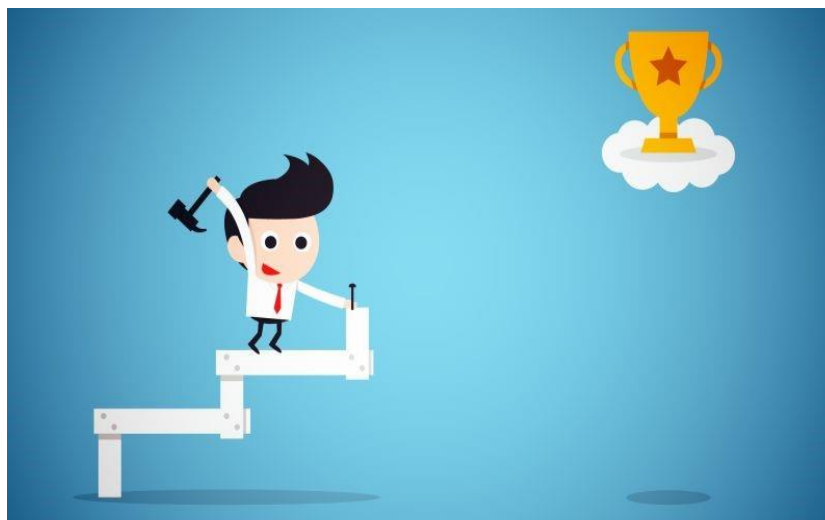
Register Strategy (social statistics)



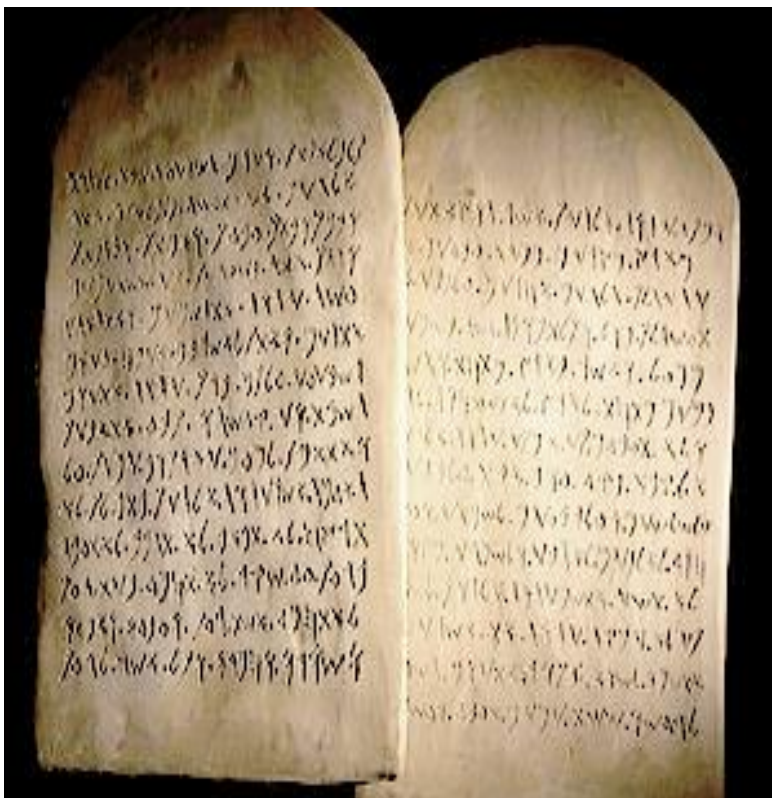
Microdata are cataloged and exchanged as module data

Module data is characterized by being

- Coherent (concepts and variables),
- Stable,
- Documented and
- Safe



Module data – the 9 commandments



1. Module data comprise all person-identifiable data exchanged including sample populations.
2. Publications must be based on module data.
3. Module data must be published at the same time as NYT and Statbank.
4. Internal and external data must be delivered as module data.

The 9 commandments...



5. The key for persons must be the unique defined [*person-ident*] which is generated for statistical purposes.
6. Module data must be documented in TIMES.
7. A test-report should be prepared for module data; even if a new version of data is made.
8. No module data are erased.
9. Dispensation may be granted by following a formal requesting procedure.

Documentation of **module data**

In addition to general requirements for documentation a 'test report' must be prepared for every version of a microdataset. The 'test report'

- ☐ Describes the quality of every version of data
- ☐ Is sent to users of the data who can subscribe on the series of data
- ☐ is documentation for all changes in data which is considered as relevant for the users.

Module data - benefits

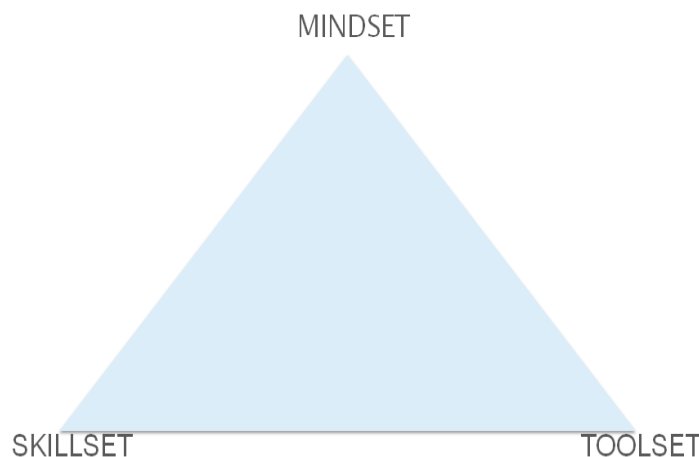


- Standardized basis of data provides data consistency and improved quality
- Systematic knowledge sharing and a culture of collaboration
- Data is displayed which provides transparency
- Published and exchanged microdata is preserved and documented
- The individual data owners (subject units) save time, making data accessible to all users only once
- Users are granted admittance to specific data series only once
- Combining series of data is easier when data/documentation is transparent
- Users have quick access to data as they collect data themselves

Module data - challenges



- A system is needed which can handle the required functionalities
- Maintenance is extremely important
- Resources are felt to be scarce and focus is on producing statistical figures
- The need for documentation can be difficult to understand
- Change of mindset/skillset/toolset



Administrative framework – Front page

Secretariat

General info

- Front page
- Intranet page
- Rules
- Guidelines
- Tips and tricks

Overview

- Search function

Create MD

- Form for creating area
- Step by step

Apply for access to data

- Instructions
- Application form

Module data

- Expected publications
- Data
- Create test report
- Send information mail

**DANMARKS
STATISTIK**

**PSD
Administrationssystem**

Sekretariat

Generelt
Startside
PSD's Intranetside
De 9 bud
Guidelines
Tips & Tricks

Oversigt - tildelte adgange
Søg udfra dataserie
Søg udfra ident

Opret MD
Blanket - Opret område
Sådan gør du - trin for trin

Søg adgang til data
Vejledning
Ansøgningsblanket

Moduldata

Forventede offentliggørelser
Dataserier
Dan e-testrapport
Send Infomail

Startside

PSD Administrationssystemet giver dig et overblik over, hvordan du lægger data ud, hvilke data der allerede ligger tilgængelige, og ikke mindst hvordan du får adgang til at benytte data!

Har du brug for yderligere hjælp, kan du henvende dig til sekretariatet, som består af:

Mikkel Zimmermann
Chefkonsulent
Tlf. 3917 3043
mzi@dst.dk

Margit London
Afdelingsleder
Tlf. 3917 3932
mlo@dst.dk

**For help – please
contact**

Lidt praktisk info i forbindelse med brug af systemet

Når et link aktiveres, bør man opdatere den kaldte side med tryk på (F5) for at være sikker på, at den seneste version vises. Dette skyldes, at Internet Explorer af hensyn til hastighed gemmer siden fra sidst, man aktiverede den.

Navigation frem og tilbage mellem siderne foregår vha. de store pile øverst til venstre.

Siden tager sig bedst ud (afrundede hjørner mm), hvis man sætter sin Explorer op uden kompatibilitetsvisning. Dette gøres ved i Explorer at vælge 'Funktioner' og 'Indstillinger for kompatibilitetsvisning' og fjerne de 3 nederste flueben.

Eller tryk F12 og vælg "Browsertilstand: Internet Explorer 9" allerøverst og næstsidste valgmulighed i værktøjslinjen.

Administrative framework – Organisation of data

The screenshot shows the PSD Administrationssystem interface. The top header features the 'DANMARKS STATISTIK' logo on the left and the 'PSD Administrationssystem' title in the center. A left sidebar contains navigation links under 'Sekretariat', 'Generelt', 'Oversigt - tildelte adgange', and 'Opret MD'. The main content area is titled 'Databases' and contains an overview of all available data series, a list of explanations, and two dropdown menus for selecting a topic and area.

**DANMARKS
STATISTIK**

**PSD
Administrationssystem**

Sekretariat

Databases

Oversigt over samtlige tilgængelige dataserier fordelt på div. emneområder

Forklaringer

- 1) Metadokumentation - Findes i Times4.
- 2) Testrapport - fremkommer ved klik på serienavnet.
- 3) Tabelbeskrivelse/datastruktur - fremkommer ved klik på datasætnavnet.
- 4) Infomails - Se udsendte informationer ved klik på "Vis Infomails"
- 5) Opdateringsstatus - angiver om datasættet stadig opdateres. Når intet er angivet, er den stadig aktiv.

Vælg emne ▼ Vælg område ▼

Follow the structure for publishing and documentation

Administrative framework – Overview of data series

Sekretariat

Generelt

Startside
PSD's Intranetside
De 9 bud
Guidelines
Tips & Tricks

Oversigt - tildelte adgange

Søg ud fra dataserie
Søg ud fra ident

Opret MD

Blanket - Opret område
Sådan gør du - trin for trin

Søg adgang til data

Vejledning
Ansøgningsblanket

Moduldata

Forventede offentliggørelser
Dataserier
Dan e-testrapport
Send Infomail

Dataserier

Oversigt over samtlige tilgængelige dataserier fordelt på div. emneområder

Forklaringer

- 1) Metadokumentation - Findes i Times4.
- 2) Testrapport - fremkommer ved klik på serienavnet.
- 3) Tabelbeskrivelse/datastruktur - fremkommer ved klik på datasætnavnet.
- 4) Infomails - Se udsendte informationer ved klik på "Vis Infomails"
- 5) Opdateringsstatus - angiver om datasættet stadig opdateres. Når intet er angivet, er den stadig aktiv.

Uddannelse og Viden ▼

Befolkningens uddannelsesst. ▼

Serienavn

Højeste fuldførte uddannelse
Højeste fuldførte uddannelse - 2014
Indvandrernes medbragte uddannelse
Uddannelses-status

Datasætnavn

D400600.PSD_HOEJSTE_FULDFOERT_UDD
D400600.PSD_HFUDD
D400600.PSD_INDVANDRERNES_MEDBR_UDD
D400600.PSD_UDD_STATUS


Opdateringsstatus

Afsluttet

Udsendt Info

Vis Infomails
Vis Infomails
Vis Infomails
Vis Infomails

Administrative framework – Overview of datasets


**DANMARKS
STATISTIK**

**PSD
Administrationssystem**

Sekretariat

Testrapport

Serie: Uddannelses-status
Datasæt: D400600.PSD_UDD_STATUS

<u>Referencetid/-periode</u>	<u>Versionsnr</u>	<u>Godkendelse</u>	<u>Ansvarlig</u>	<u>Fejlmeldelse</u>
31-12-2016	1	Testrapport godkendt	SMS	Vis testrapport
31-12-2015	2	Testrapport godkendt	SMS	Vis testrapport
31-12-2015	1	Testrapport godkendt	SMS	Vis testrapport
31-12-2014	2	Testrapport godkendt	SMS	Vis testrapport
31-12-2014	1	Testrapport godkendt	SMS	Vis testrapport
31-12-2013	2	Testrapport godkendt	SMS	Vis testrapport
31-12-2013	1	Testrapport godkendt.	SMS	Vis testrapport
31-12-2012	2	Testrapport godkendt	SMS	Vis testrapport
31-12-2012	1	Testrapport godkendt.	SMS	Vis testrapport
31-12-2011	2	Testrapport godkendt.	SMS	Vis testrapport
31-12-2011	1	Testrapport godkendt.	SMS	Vis testrapport
31-12-2010	2	Testrapport godkendt.	SMS	Vis testrapport
31-12-2010	1	Testrapport godkendt.	SMS	Vis testrapport
31-12-2009	2	Testrapport godkendt.	SMS	Vis testrapport
31-12-2009	1	Testrapport godkendt.	SMS	Vis testrapport
31-12-2008	3	Testrapport godkendt.	SMS	Vis testrapport

Moduldata

Generelt

- Startside
- PSD's Intranetside
- De 9 bud
- Guidelines
- Tips & Tricks

Oversigt - tildelte adgange

- Søg ud fra dataserie
- Søg ud fra ident

Opret MD

- Blanket - Opret område
- Sådan gør du - trin for trin

Søg adgang til data

- Vejledning
- Ansøgningsblanket

Forventede offentliggørelser

- Dataserier
- Dan e-testrapport
- Send Infomail

Reference time

Vers


Status for test report

Responsible for the statistic

Show test report

Users are automatically notified by mail when new data is available

Administrative framework – Overview of variables

**DANMARKS
STATISTIK**

Sekretariat

Generelt
Startside
PSD's Intranetside
De 9 bud
Guidelines
Tips & Tricks

Oversigt - tildelte adgange
Søg ud fra dataserie
Søg ud fra ident

Opret MD
Blanket - Opret område
Sådan gør du - trin for trin

Søg adgang til data
Vejledning
Ansøgningsblanket

Moduldata

Forventede offentliggørelser
Dataserier
Dan e-testrapport
Send Infomail

**PSD
Administrationssystem**

Tabelbeskrivelse

Serie: Uddannelses-status

Datasæt: D400600.PSD_UDD_STATUS


- Højeste fuldførte almenuddannelse - Højeste fuldførte erhvervskompetencegivende uddannelse - Højeste fuldførte uddannelse, uanset om den er almen eller erhvervskompetencegivende - Igangværende uddannelse - for personer der er i uddannelse pr. 1. oktober. samt institutionsnummer og fuldførelsestidspunkt på den højst fuldført uddannelse, højst fuldførte almene uddannelse, højst fuldførte erhvervskompetencegivende uddannelse og seneste fuldførte uddannelse. Bemærk, at variablen HF-KILDE i dette datasæt angiver kilden til den højst fuldførte uddannelse.

Kolonnenavn	Datatype	Længde	Kommentar
ALMAUDD	NUMBER	4	Kode for højst fuldført almen uddannelse pr. 1. oktober Den højeste fuldførte almen uddannelse er, som det vigtigste kriterium, fastlagt i forhold til hovedgrupperne i den danske uddannelsesnomenk
ALMINSTNR	NUMBER	6	Institutionsnummer på den institution, hvor højst fuldført almene uddannelse er afsluttet. Se under variablen INSTNR for at få yderligere dokumentation for institutionsnummeret.
ALM_VFRA	DATE	7	Tidspunkt for opnået højst fuldførte almenuddannelse.
DATOPRET	DATE	7	Dato for oprettelse af moduldata
ERHAUDD	NUMBER	4	Kode for højst fuldførte erhvervskompetencegivende uddannelse. Den højeste fuldførte erhvervskompetencegivende uddannelse er, som det vigtigste kriterium, fastlagt i forhold til hovedgrupperne i de
ERHINSTNR	NUMBER	6	Institutionsnummer på den institution, hvor højst fuldført erhvervskompetencegivende uddannelse er afsluttet. Se under variablen INSTNR for at få yderligere dokumentation for institutionsnummeret.
ERH_VFRA	DATE	7	Tidspunkt for opnået højst fuldført erhvervskompetencegivende uddannelse
HFAUDD	NUMBER	4	Kode for højst fuldført uddannelse pr. 1. oktober Den højeste fuldførte uddannelse er, som det vigtigste kriterium, fastlagt i forhold til hovedgrupperne i den danske uddannelsesnomenklatur

Name	Type	Length	Explanation and
------	------	--------	-----------------

Administrative framework – Creation of test reports

http://srvblaise2.dst.local/?PopupBlocked=false&js=yes&sh=934&sw=1270 - Moduldata, testrapporte - Internet Explorer

 **Moduldata, testrapporter**

Testrapport til kvalitetssikring af Moduldata(MD)

Kontonummer:

KontorNavn:

Testrapporten er oprettet af (ident):

Dato for oprettelse:

Version:

Emnegruppe:

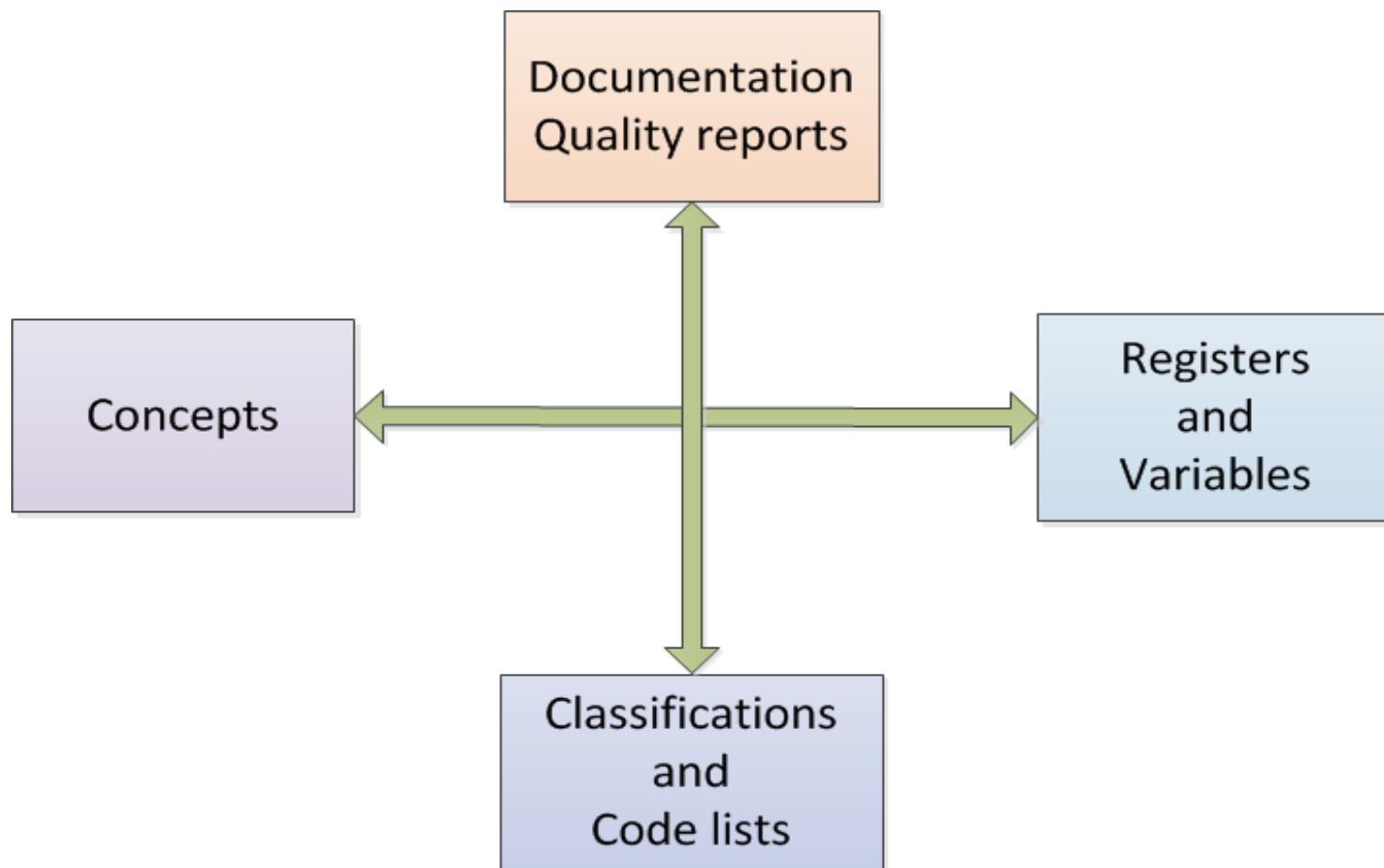
Benyttes MD til offentliggørelser?

Administrative framework – content of test report

Test Report <i>Education and Knowledge - Educational status</i> <i>per. 31-12-2016</i>	
Office	1. Population and education
Subject Area	Education and Knowledge
Series Name	Educational status
The test report was created by (ident)	sms
Date of creation of report	27th of June 2017
Publishing Frequency	Annual Statistics
Version	1
Reason for the new version (version> 1)	
How is MD calculated?	<input checked="" type="checkbox"/> Reference Date: 31-12-2016 <input type="checkbox"/> Reference period:
Single events or longitudinal data?	<input checked="" type="checkbox"/> Single events during the period <input type="checkbox"/> Longitudinal data
Census or sampling?	<input checked="" type="checkbox"/> census <input type="checkbox"/> Randomized counting
Matrix reference in StatBank	HFUDD10, HFUDD15, HFUDD20
Is MD used for publications?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No
Is the publication directly based on MD? (If 'No' indicate explanation)	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No explanation:
Date of when data is available as MD. If data is published, specify the date of publication.	21st of June 2017
Is there any correspondence between publication and MD?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

For complete test report – please consult **handout**

Metadata model / Diamond model



Presentation of metadata-portal in Colectica: <http://colecticapub:8081/>

Drivers to keep the guidelines and policies in mind



Drivers to keep the guidelines and policies in mind...

- Visible and available contact person
- Yearly brush-up and briefing meetings with statistic responsables
- Newsletters are sent to stakeholders
- Short workshops or courses
- Frequently reminders of the importance of obeying the rules because this
 - Increases the quality of data
 - Reduces the risk of errors
 - Saves time (in the long term) when the correct data and documentation is displayed and delivered the first time

Presentation of data catalogue and documentation in Research Services Denmark

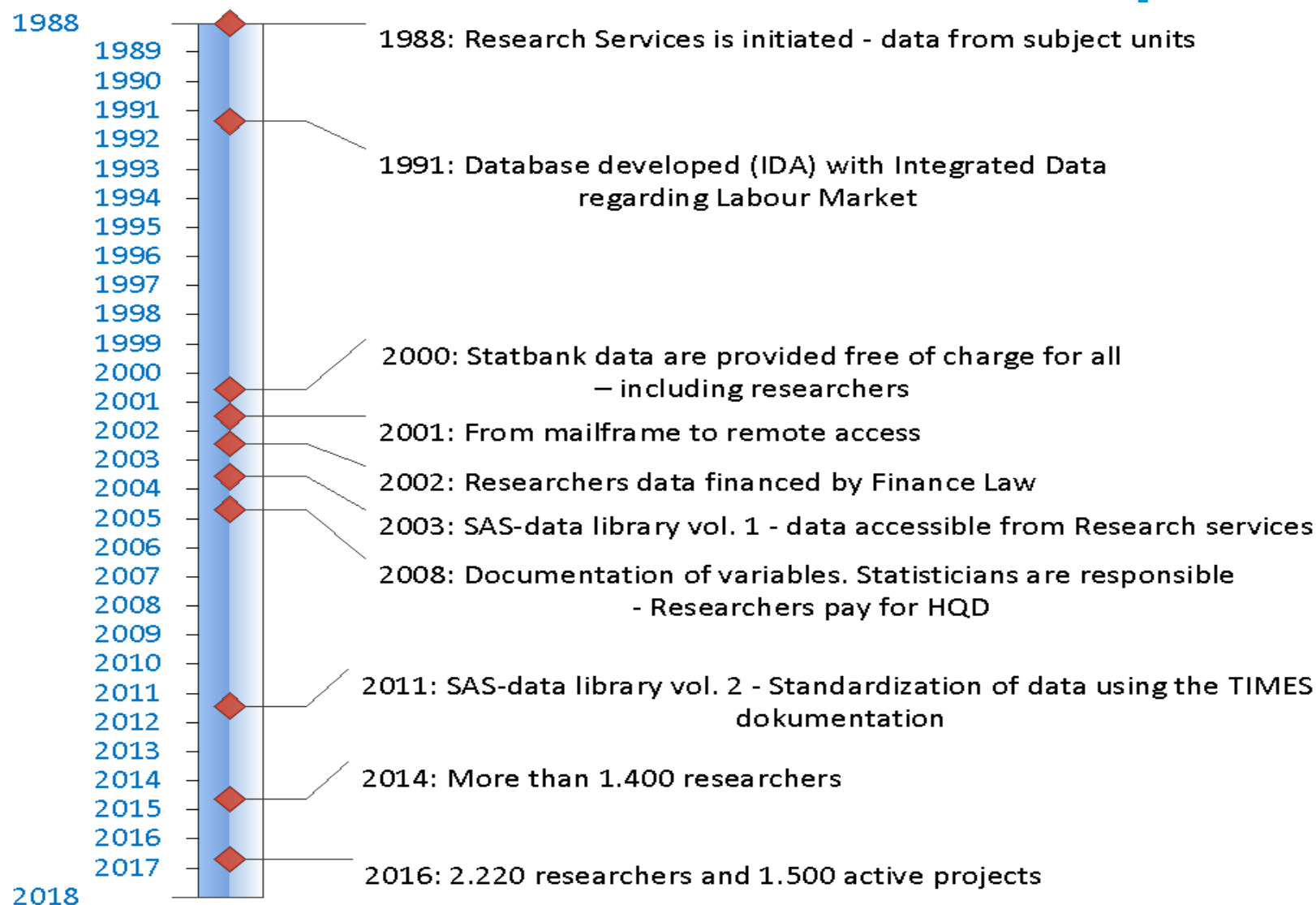


Content of the presentation

- **Background**
- **Data Catalogue**
 - Content and Purpose
 - Policy
 - Datastructure - Standardization and harmonization
 - SAS-data Library presented as a data catalogue
 - Data catalogue: presentation with links to documentation
 - Data catalogue: maintenance and organisation

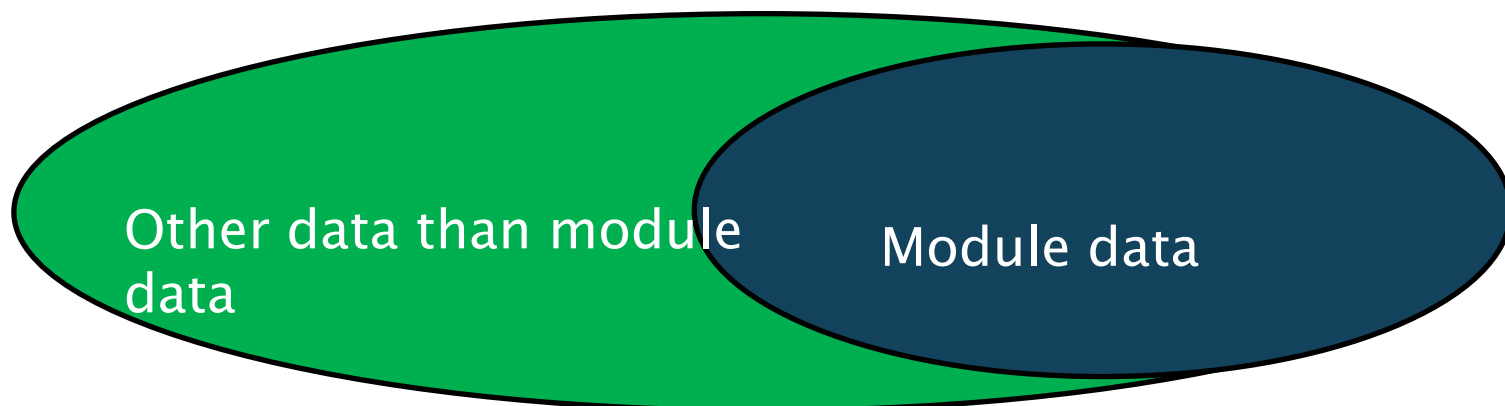


Research Services – historical retrospective



Data catalogue in Research Services (RS) 2018

- App. 280 datasets with more than 12.000 variables in the data catalogue



97 module datasets are available in 2018

1:1 relationship between RS' data catalogue and module data series



Data catalogue in Research Services (RS) 2018



- Purpose of data catalogue
 - To make micro data from Statistics Denmark available for researchers on a standardized form
- Content of data catalogue
 - All data from the statistical program (published data (Module data))
 - Unpublished register data
 - Data from government agencies
 - Special deliveries
 - Historical data / longitudinal data)

Policy – principles for data in the catalogue

- ❑ Approval from the owner of data (Subject units, external owner – Research Services do *not* own data)
- ❑ Uniform naming and structure of data over time
- ❑ Micro data with key variables to link individuals, enterprises, properties, etc. – direct identifiers
- ❑ Continuous (annually, quarterly, monthly) delivery of data over a certain period



Policy



- Accessibility of data:

- High priority data: 1 week from publication / release
- Low priority data: 4 weeks or on demand

- Accuracy of data:

- The content of a register in the catalogue resembles only published data (content, names etc.)

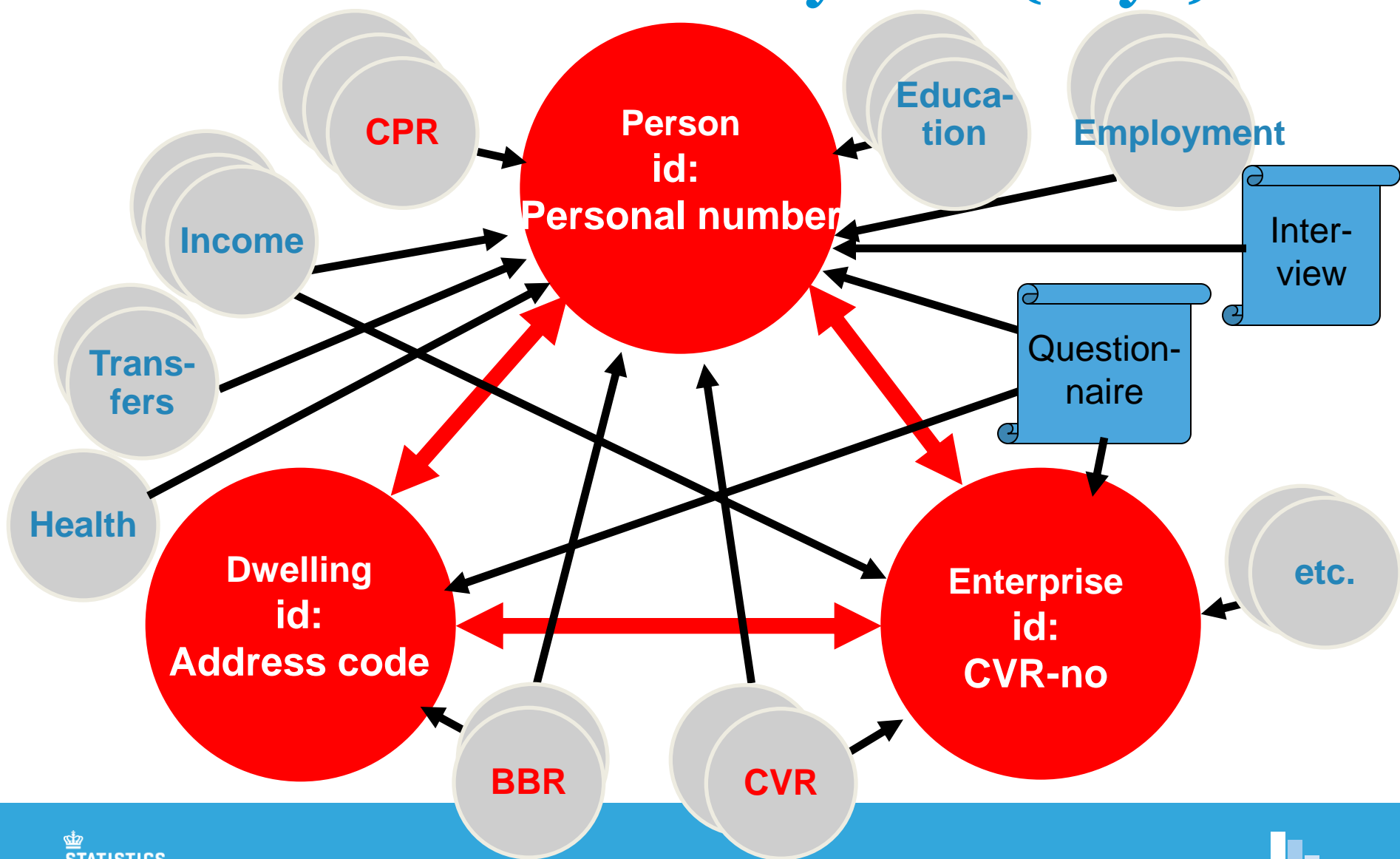
- Data confidentiality

- Encryption of key variables
- Only for researchers use - not for administrative use

Data structure: harmonization and documentation

- **Naming of datasets** consist of 3-8 signs followed by the long name;
 - UDDA – Education,
 - RAS - Labourforce statistics
- **2011 SAS-data catalogue vol. 2**
 - Standardization of data using the Times documentation
- **Variable names** adopted from TIMES;
 - Identical content over time – identical name over time, shared variable description
 - Ownership of the documentation by the subjects units
- **Variable structure**
 - Numerical/character over time
- **Key variables** (research services)
 - Name (identical across datasets and time)
 - Format (direct identifiers to be de-identified before delivered to research projects)
 - Documentation is made by RS and not by subject units

The overall statistical system (keys)



Data presented as a data catalogue

Grunddata
 Afidentificeringsliste
 Emneopdelt registeroversigt
 Liste med samtlige variable
 Sorteringsvariable
 Problematiske afidentificeringsvariable i grundregistre

Register	Register titel	Forlebsdata	Referencetype	Opdateringsfrekvens	Særlige forhold	Første år	Seneste år	Antal dataset	Seneste referencedata	Seneste opdateringsdato	Næste opdateringsdato (foreløbig)	Ansvarlig indhold	Ansvarlig opdatering
ADOP	Adoptioner fra CPR		PERIODE	ÅR		1988	2009	22	31/12/2009	31/12/2010	31/12/2010	KAE	MTL
AEFB	Ældre Dokumentation forebyggende hjemmebesøg		PERIODE	ÅR		2008	2015	8	31/12/2015	14/06/2016	01/07/2017	CLR	MTL
AEFV	Ældre Dokumentation visiteret hjemmehjælp til valø		PERIODE	ÅR		2008	2015	8	31/12/2015	14/06/2016	01/07/2017	CLR	MTL
AEH	Ældre Dokumentation leveret hjemmehjælp		PERIODE	ÅR		2011	2015	5	31/12/2015	14/06/2016	01/07/2017	CLR	MTL
AEPR	Ældre Dokumentation modtagere af hjemmehjælp plejehol		PERIODE	ÅR		2008	2015	8	31/12/2015	14/06/2016	01/07/2017	CLR	MTL
AEPI	Ældre Dokumentation borgere på plejehjem bolig (imputeret)		PERIODE	ÅR		2008	2015	8	31/12/2015	04/11/2016	31/01/2017	CLR	NN
AETR	Ældre Dokumentation genoptræning vedligeholdelsestræning		PERIODE	ÅR		2008	2015	8	31/12/2015	14/06/2016	01/07/2017	CLR	MTL
AKAS	Akasser medlemmer		STATUS	ÅR		2008	2014	7	31/12/2014	31/05/2016	01/06/2017	KBU	MTL
AKM	Arbejdsklassifikationsmodul		PERIODE	ÅR		1976	2014	39	31/12/2014	25/02/2016	01/02/2017	CLR	CAG
AMFQ	Arbejdsmarkedet foranstaltninger (AMFORA)		PERIODE	ÅR		1994	2006	13	31/12/2006	27/09/2011	27/09/2011	KBU	NN
AMRTN	Arbejdsmarkedet for ansættelse med tilsætnormering	JA	PERIODE	ÅR		2008	2014	1	31/12/2014	30/10/2016	01/03/2017	CLR	CAG
AMRUN	Arbejdsmarkedet for ansættelse uden tilsætnormering		PERIODE	ÅR		2008	2014	7	31/12/2014	24/01/2017	01/02/2017	CLR	CAG
ARB	Arbejdssteder fra ESR (foreløbig tom 16 dec 2015)		PERIODE	ÅR		1999	2015	17	31/12/2015	01/03/2016	20/01/2017	SVP	JNI
ATAA	Anmeldte arbejdsulykker	JA	PERIODE	ÅR	JA	2005	2014	1	31/12/2014	24/02/2016	31/01/2017	KBU	NRA
ATRT	Arbejdsulykkes risikobaseret tilsyn	JA	PERIODE	ÅR	JA	2012	2015	2	31/12/2015	09/12/2016	30/06/2017	KBU	NN
BARNFORA	Hervisninger mellem barn og forældre		ÅR		JA	1980	2015	1	31/12/2015	14/06/2016	14/06/2016	KAE	NRA
BBSB	Bvope og bolloregister (BED) - byvninger		STATUS	ÅR	JA	2005	2016	12	31/12/2015	05/07/2016	30/04/2017	RBN	MTL
BBSB	Bvope og bolloregister (BED) - enheder		STATUS	ÅR		2005	2016	12	31/12/2015	05/07/2016	30/04/2017	RBN	MTL
BG	Beskæftigede CFAM		PERIODE	ÅR		1985	2007	23	31/12/2007	27/09/2011	27/09/2011	KBU	NN
BEE	Befolkningen (år)		STATUS	ÅR		1986	2016	31	31/12/2015	11/02/2016	23/02/2017	MNH	MTL
BEFADIS	Adresseregister	JA	PERIODE	ÅR		2015	2015	1	31/12/2015	01/04/2016	28/02/2017	MNH	MTL
BEFBOP	Bopælændringer	JA	PERIODE	ÅR		2015	2015	1	31/12/2015	01/04/2016	28/02/2017	MNH	MTL
BEFK	Befolkningen (kv.)	JA	STATUS	KV/T		2008	2016	9	30/09/2016	09/11/2016	09/02/2017	MNH	MTL
BEL	Detaljeret levmøddagerdata fra e-indkomst tom 2015		PERIODE	KV/T	JA	2008	2015	8	31/12/2015	09/01/2017	09/04/2017	CGM	CAG
BELBR	Aggrigerede levmøddagerdata på brancher (127) tom 2016 kv2_2		PERIODE	KV/T	JA	2008	2016	9	30/06/2016	09/01/2017	09/04/2017	CGM	CAG
BELBRTI	Aggrigerede levmøddagerdata på brancher (10) tom 2016 kv2_2		PERIODE	KV/T	JA	2008	2016	9	30/06/2016	09/01/2017	09/04/2017	CGM	CAG

- Abbreviated register name
- Register name
- First and last year
- Frequency of update
- Date for next update

Data presented as a data catalogue

- Ex. Population

BEF - Befolkningen (år)			År																															
TIDSPER	Højniveaustokumentation	Label	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
ABST_ID	ABST_ID	Abgørelse ID	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
ALDER	ALDER	Ålder pr. 1. januar	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
BOPKOM	BOPKOM	Adresse(n) (vej, husnummer, huslejetilskuds- etape, sidelejemnummer)	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
BOP_VPSA	BOP_VPSA	Data for boplysningsindvanding	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
CVST	CVST	Civilstand	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
CV_VPSA	CV_VPSA	Civilstandsdata	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
CVTJEB	CVTJEB	CVRtype	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
CVRTYPE	CVRTYPE	CVRtype	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
EPALLE	EPALLE	Personnummer E-falte	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FAMBL_ID	FAMBL_ID	Familien identificerende nummer	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FAMBL_TYPE	FAMBL_TYPE	Familietype	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FAR_ID	FAR_ID	Faderens Person ID	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FAR_K	FAR_K	Far(K)	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FM_MARK	FM_MARK	Fødselsmærkning	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FOEDRES_HOGE	FOEDRES_HOGE	Kode for personens fødselsregistreringssted	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FOED_DAG	FOED_DAG	Fødselsdato	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
FOERSTE_INDVANDRING	FOERSTE_INDVANDRING	Første indvandringsdato	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GENERATION	GENERATION	Kategoriisering af 2. og 3. generationsindvandrere (efter DS-definitio	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
HUSTYPE	HUSTYPE	Hustandstype	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
ILTYPE	ILTYPE	Indvandrers efterkommere - personer med dansk opvækst	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
KON	KON	Køn	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
KOM	KOM	Kommunekode	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
MOR_ID	MOR_ID	Moders Person ID	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
OPKOM	OPKOM	Opgangsadresse i kommunen	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
OPV_LAND	OPV_LAND	Opvækstland	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
PLAGE	PLAGE	Familiestatus	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
PNR	PNR	Personnummer	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
REG	REG	Region	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
SENESTE_INDVANDRING	SENESTE_INDVANDRING	Seneste indvandringsdato	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
STATSGR	STATSGR	Statsoptegnelse	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	

Data catalogue: presentation with links to documentation

- The catalogue is maintained as an interactive catalogue with hyperlinks to the documentation at register *and* variable level (not 1:1-documentation at register-level)
- The catalogue is updated every night.
- Formats and documentation is available on the researchers servers

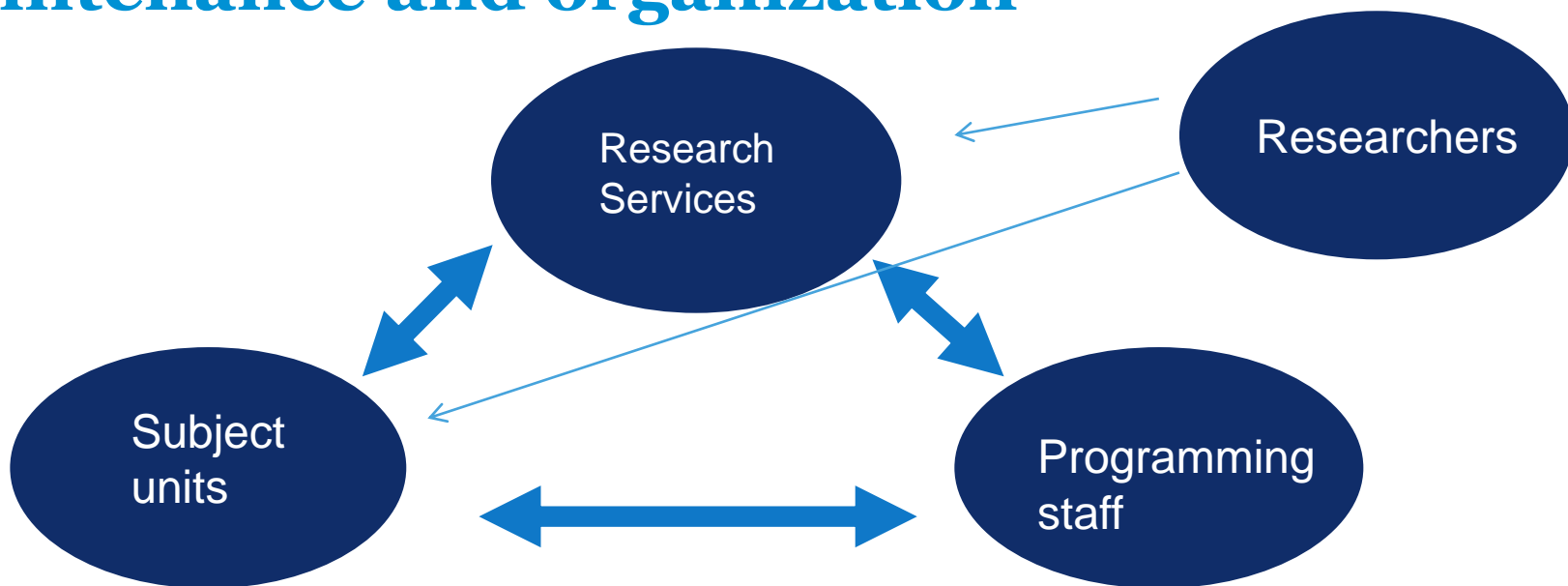


https://www.dst.dk/da/TilSalg/Forskningservice/Data/Register_Variabeloversigter

Data catalogue: maintenance and organisation

- Data catalogue team in RS
 - Is responsible that decisions about the data catalogue comply with the data policy
 - Is responsible for the maintenance of up-dates of the catalogue based on FSE-meta every night
 - Handle day-to-day questions and decisions about data catalogue
- Service managers in RS
 - Have basic knowledge about the content of the registers
 - Cooperate with Subject Units
 - Only have reading access and not writing access to data catalogue
- Programming staff
 - Has the technical responsibility - data management (maintenance, updates)
 - Is responsible for the standardization of the content of the data catalogue (every register is updated by using the same SAS-template which provides the standardization and harmonization of the data)

Data-catalogue: maintenance and organization

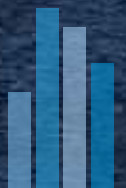


Corporation between researchers, subject units and Research Services/programming staff

Mostly used registers in Denmark

Registers	Long name	Number of extracts in 2017	
BEF	Population	6646	
UDDA	Education	6312	
IND	Personal income	5932	
AKM	Working classifications module	4187	
LPRDIAG	Hospitalized patient – diagnoses	3723	
RAS	Register–Based Labour Force Statistics	3261	
LMDB	Prescription Medicine data base	2768	
IDAP	Integrated data base on labour market research – social data	2137	
IDAN	Integrated data base on labour market research – salary data	1900	
FAM	Family income	1725	

High Quality Documentation – (HQD)

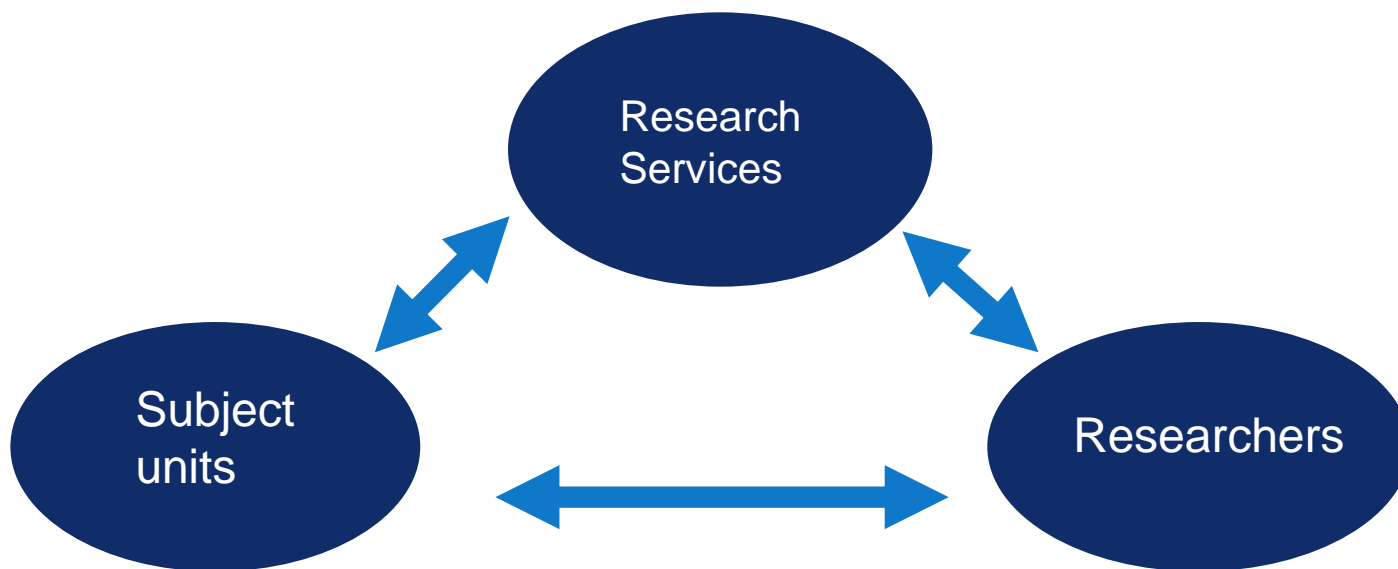


Historical background

- **Researchers demand for better documentation caused development of High quality documentation (HQD)**
- **Researchers familiarity with TIMES-documentation system – expansion of existing documentation with new elements (historic, data breach etc.)**
- **Financing – development/expansion of TIMES to HQD and every year since then**
- **690 variables documented since 2007/2008**



The process of producing the documentation



- **Corporation between researchers, subject units and Research Services (one year process)**
- **Final approvement by the steering committee. Payment is done at this time.**

Special template - examples



- **Systematic Documentation:**
 - Detailed variable description, including data breaches
 - Validity (period) /population
 - Tables and graphs
 - Formats
 - Appendices with special interest

<http://www.dst.dk/da/TilSalg/Forskningservice/Dokumentation/hoejkvalitetsvariable.aspx>



Areas with special conditions

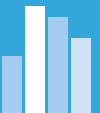


- Employee statistics:
 - Detailed data is made available one year after the reference period
 - App. 5 months after the reference quarter detailed personal data are available but the associated workplace variables are aggregated
- Drug data – Danish Medicines Agency must approve
 - The research project
 - Access by the individual researcher
 - Private sector's admittance to drug data is even more restrictive

Benefits of having a well-documented data catalogue

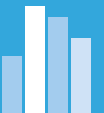


- Standardized basis of data provides data consistency – makes access for the researchers more easy
- Data is displayed which provides transparency for the researchers
- The individual data owners (subject units) save time, making data accessible to all users only once
- The individual data owners (subject units) save time, making documentation only once and by making it accessible to all users on the net
- Combining series of data is easier when data and documentation is transparent



Challenges of having a well-documented data catalogue

- Maintenance and up-dating of the data catalogue is very important and has to be of high priority
- Allocation of resources is necessary
- The organization and the responsibility for the data catalogue must be in order



Tuesday January 30th 2018

Need of clarifications from yesterday:

1. Clarifications of which level each dataset must be assigned to
2. Interface to SIMS from the data catalogue
3. Is there a need for documentation of MUC, PUF, research room files?

POLICY suggestions...

PURPOSE OF DATA CATALOGUE:

- To make micro data from ICBS available for researchers on a standardized form.

POLICIES / PRINCIPLES for data in the catalogue:

- Contains all data from the statistical program and special deliveries
- Uniform naming and structure of variables and datasets over time
- The catalogue only displays variables and datasets where data have already been published if related to the statistical program

POLICY suggestions...

POLICIES / PRINCIPLES for data in the catalogue:

- Subject Units own the data and are responsible for documentation of datasets and variables
- Int. Relations and Stat. Coordination Department is responsible that decisions about the data in the catalogue comply with the data policy
- RS is responsible for the maintenance of updates of the catalogue
- RS handle day/to/day questions and decisions about data catalogue

Starting the work.....

- Decide on the structure of the content of the catalogue (which datasets, which variables – sufficient for all the “Needs to know”)
- Decide on the definition of a datasets and a variable
- Decide on the harmonization over time and across domains on dataset names and variable names
- Decide on which level the documentation shall be integrated in the data catalogue
- Decide on centralization of the data catalogue in RS
- Decide on maintenance and up-dating of the data catalogue in RS
- Decide on allocation of ressources to build and maintain the data catalogue
- Start with the most demanded data
- Develop an it-system to support the data catalogue



Questions?

