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

EU TWINNING PROJECT 2016-2018

ENHANCING THE QUALITY OF OFFICIAL STATISTICS

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

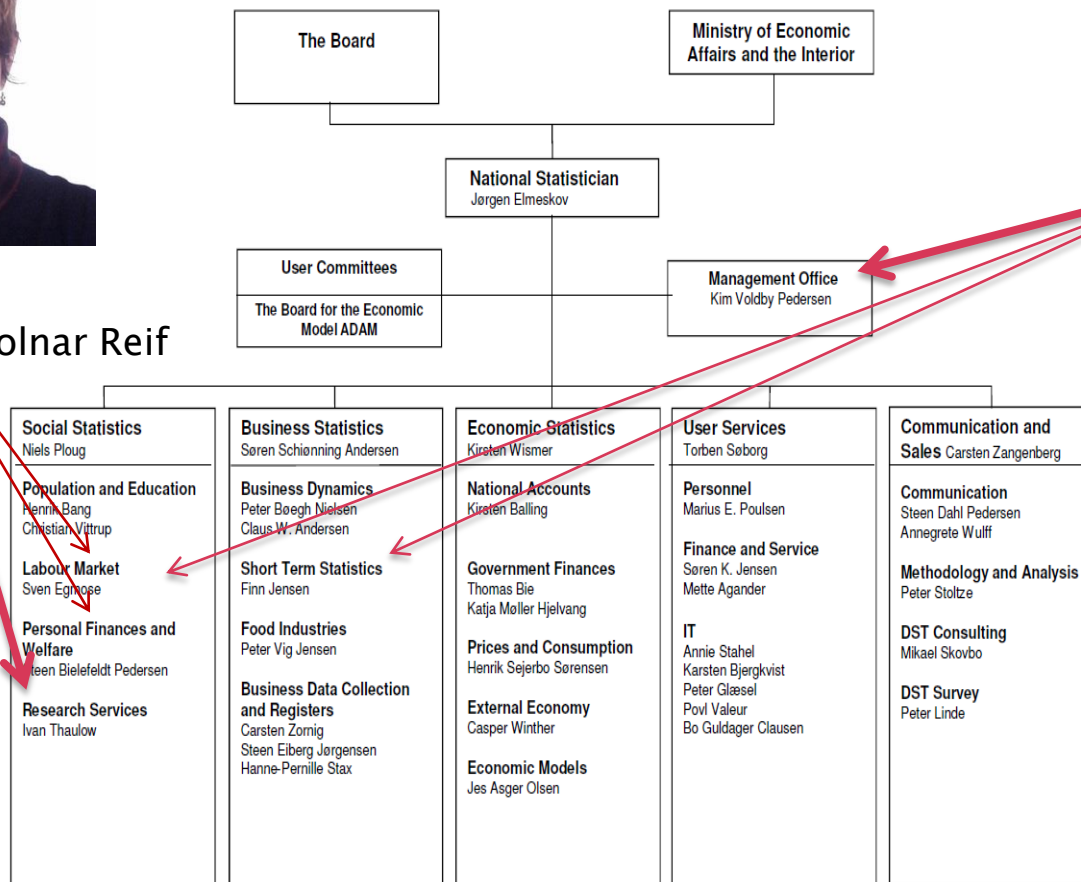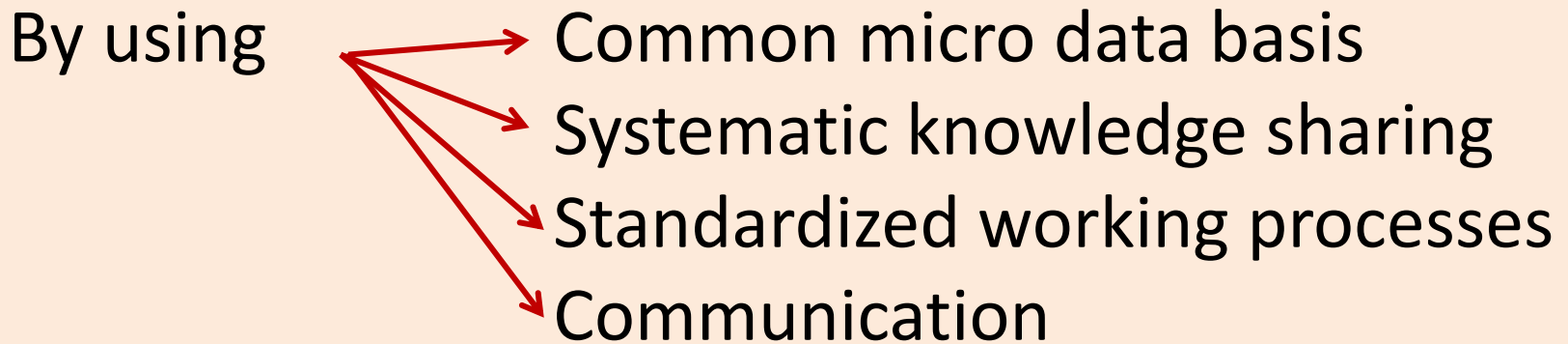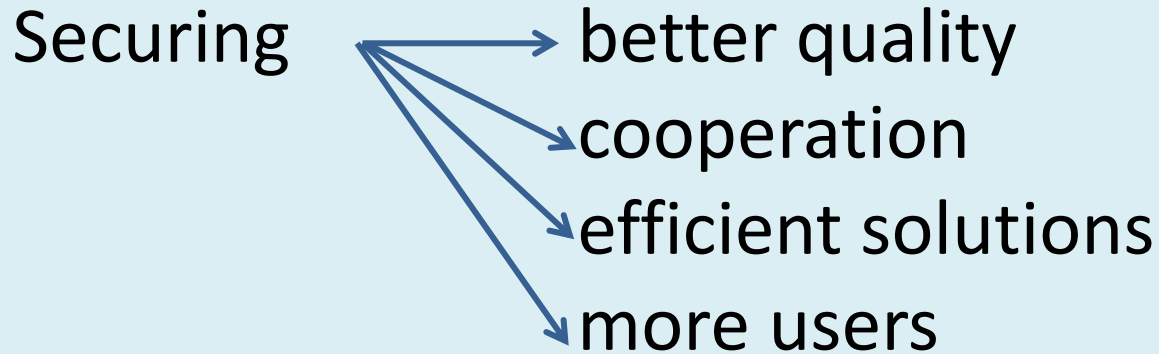|	
9:00 – 9:20	Second day review
9:20 – 14:30	**MS** and **BC**: Outlining a data catalogue policy, structure and content of data. Discussing a process for building a catalogue
14:30 – 16:00	Summary, evaluation and final remarks

Presentation by MS

Charlotte Leolnar Reif

Karin Holst Duer

Register Strategy (social statistics)

Securing → better quality
cooperation
efficient solutions
more users

By using → Common micro data basis
Systematic knowledge sharing
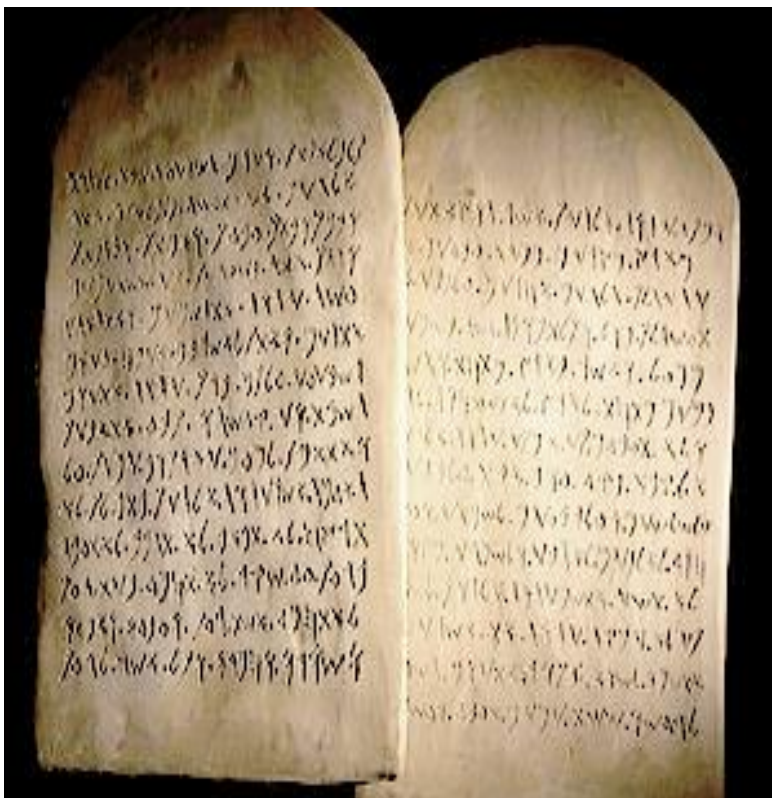Standardized working processes
Communication

Microdata are cataloged and exchanged as *module data*

Module data is characterized by being

- Coherent (concepts and variables),
- Stable,
- Documented and
- Safe
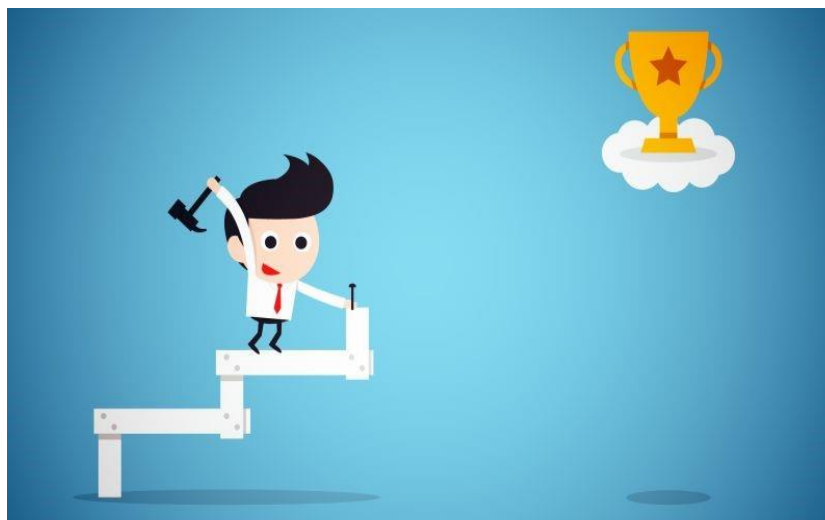
Module data – the 9 commandments

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 quickly 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

MINDSET

SKILLSET   TOOLSET

STATISTICS
DENMARK
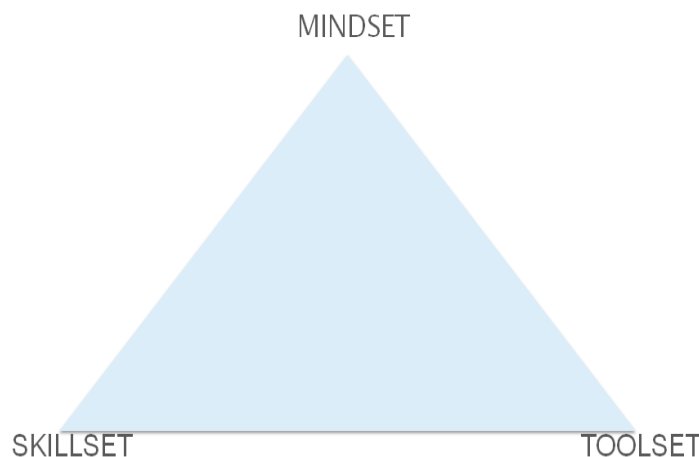
Administrative framework – Front page

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

For help – please contact

Administrative framework – Organisation of data

Follow the structure for publishing and documentation

Administrative framework – Overview of data series

Administrative framework – Overview of datasets

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

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

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 fuldføre almene uddannelse, højst fuldføre 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 fuldføre almenuddannelse.
DATOPRET	DATE	7	Dato for oprettelse af moduldata
ERHAUDD	NUMBER	4	Kode for højst fuldføre erhvervskompetencegivende uddannelse. Den højeste fuldføre 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 erhvervskompetancegivende 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

Administrative framework – content of test report

Test Report Education and Knowledge - Educational status per. 31-12-2016	
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?	☑ **Reference Date: 31-12-2016** ☐ Reference period:
Single events or longitudinal data?	☑ **Single events during the period** ☐ Longitudinal data
Census or sampling?	☑ **census** ☐ Randomized counting
Matrix reference in StatBank	HFUDD10, HFUDD15, HFUDD20
Is MD used for publications?	☑ Yes ☐ No
Is the publication directly based on MD? (If 'No' indicate explanation)	☑ Yes ☐ 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?	☑ Yes ☐ No

For complete test report – please consult handout

Metadata model / Diamond model
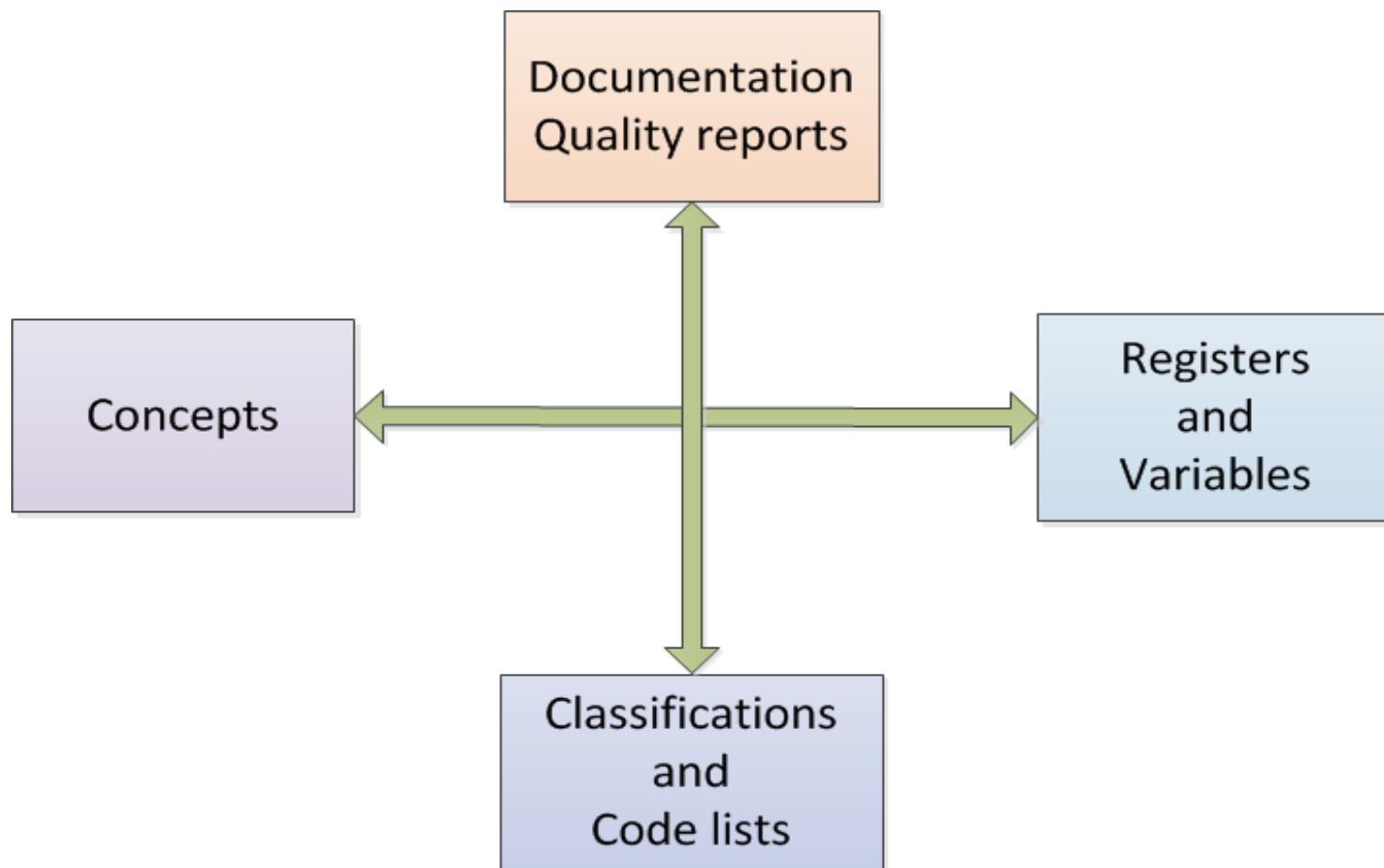
Presentation of metadata-portal in Colectica: http://colecticapub:8081/

Drivers to keep the guidelines and policies in mind

STATISTICS
DENMARK

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

- Visible and available contact person

- Yearly brush-up and briefing meetings with statistic responsibles

- 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

STATISTICS DENMARK

Presentation of data catalogue and documentation in Research Services Denmark

DANMARKS STATISTIK

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

1988: Research Services is initiated - data from subject units

1991: Database developed (IDA) with Integrated Data regarding Labour Market

2000: Statbank data are provided free of charge for all – including researchers

2001: From mailframe to remote access

2002: Researchers data financed by Finance Law

2003: SAS-data library vol. 1 - data accessible from Research services

2008: Documentation of variables. Statisticians are responsible - Researchers pay for HQD

2011: SAS-data library vol. 2 - Standardization of data using the TIMES dokumentation

2014: More than 1.400 researchers

2016: 2.220 researchers and 1.500 active projects

Data catalogue in Research Services (RS) 2018

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

Other data than module data
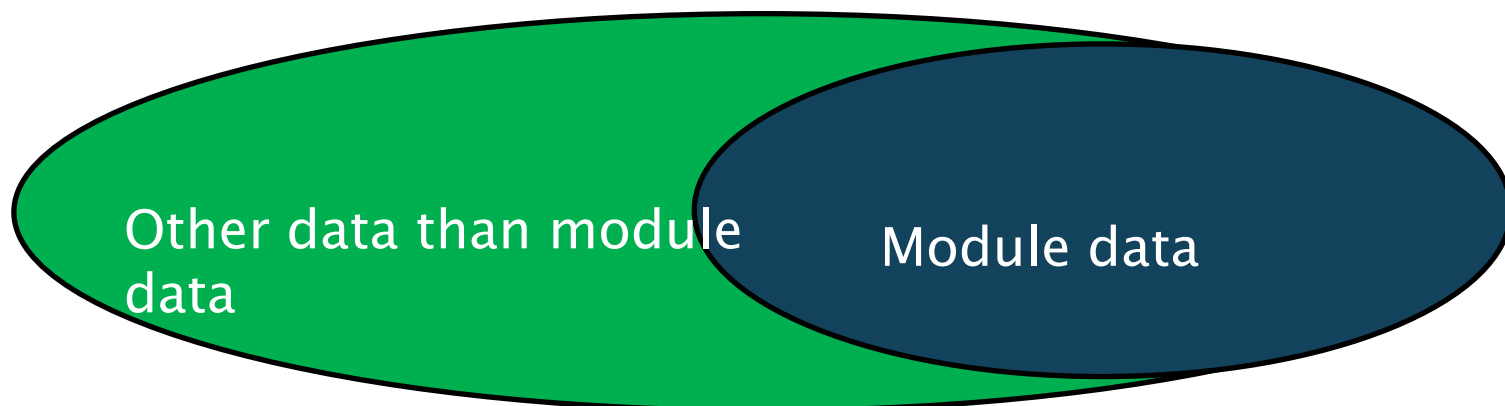
Module data

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 subjets 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

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

Data presented as a data catalogue

- Ex. Population

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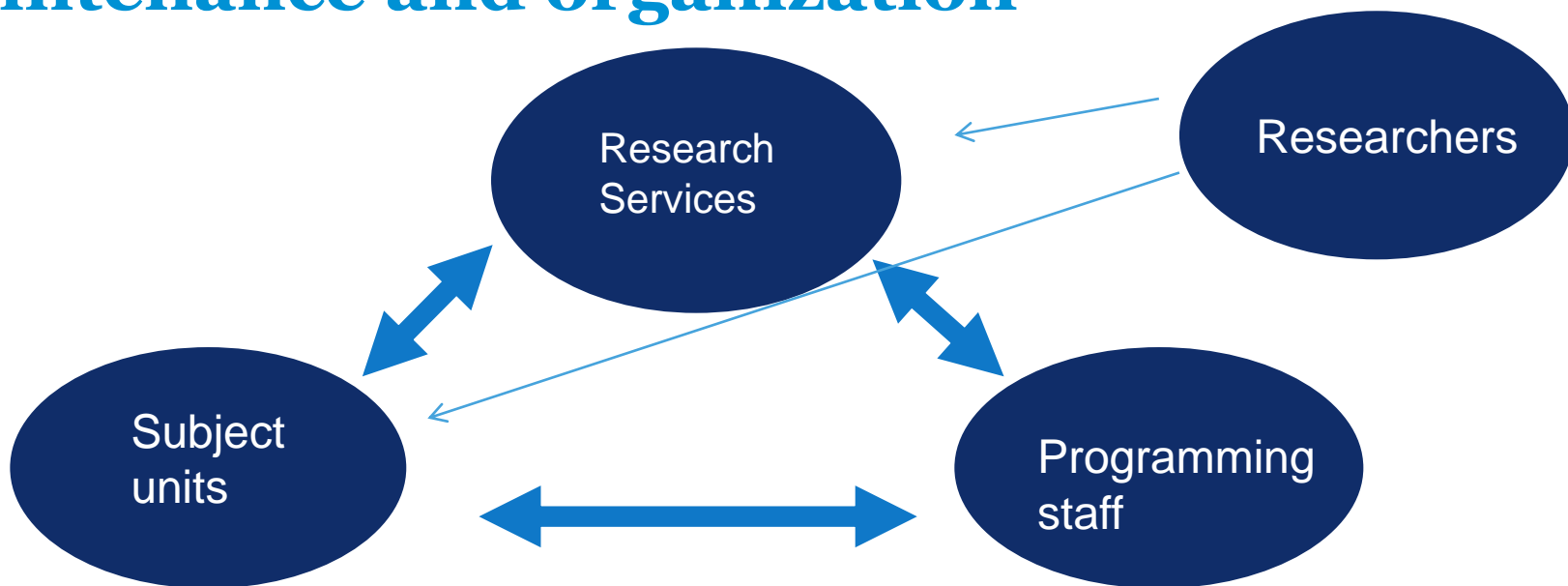
Corporation between researchers, subject units and Research Services/programming staff

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)
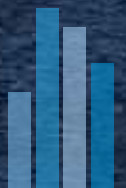
Historical background

- **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
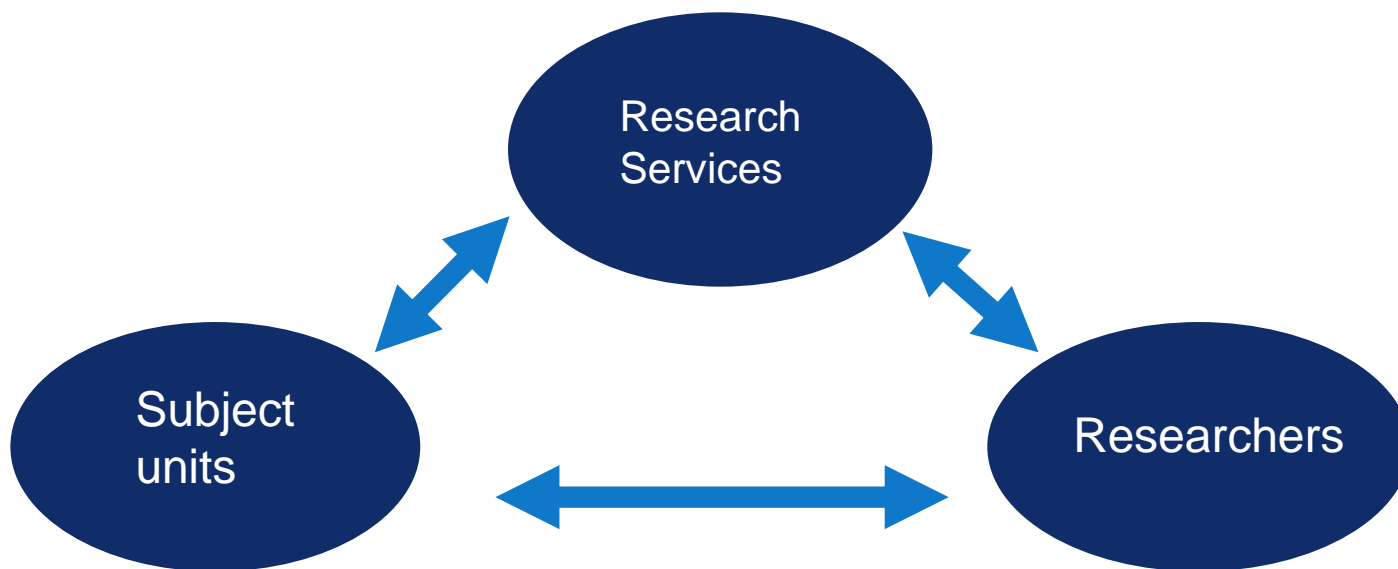
- **Corporation between researchers, subject units and Research Services (one year process)**

- **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/Forskningsservice/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
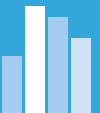
Challenges of having a well-documented data catalogue

- 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
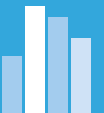
Tuesday January 30th 2018

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

Questions?