Terms of Reference

EU Twinning Project IL/12 CRIS 2015/370-467

Date: Sunday 28 January – Wednesday 31 January 2018.

Component B: Micro Data services to researchers

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

0. Mandatory results and benchmarks of the component

Mandatory results:

- *Setting up an overall program for providing researchers with access to micro-data, based on the analysis of the current situation and a designated strategic plan* (February 2018)

Benchmarks:

- **IB1:** Report and analysis of the current situation adopted (and identification of gaps to be bridged) by 2th project quarter;
- **IB2:** Strategic Plan elaborated for providing researchers with access to micro-data by 5th project quarter;
- **IB3:** Organizational and technological implementation plan, including data security, proposed by 7th project quarter;
- **IB4:** Formal organizational structure proposed to by 6th project quarter;
- **IB5:** An administrative management system adopted by 7th project quarter;
- ***IB6: Policy, structure and content of data catalogue outlined by the 8th quarter;***
- **IB7:** Guidelines for output approved by the Confidentiality Committee by 9th project quarter
- **IB8:** Proposal for legislation submitted to the Israeli Central Bureau of Statistics (ICBS) top management by 10th project quarter;

1. Purpose of the activity

The aim of the activity is to get input and discuss options for building an integrated data library, data catalogue and documentation in a professional, user-friendly and competent manner by the Israeli Central Bureau of Statistics (ICBS) with focus on policy, structure and ***content of the data catalogue***. The catalogue and associated documentation should be available both internally and externally.

The solution should be built on an infrastructure solution that supports:

i. *Public use microdata files*[1]
ii. *Scientific-use micorodata files*[2]*(MUC)*
iii. *Secure-use micorodata files*[3] *(RR)*
iv. *Databases kept by the subject units*

[1] Public Use Files (PUFs) consists of sets of records containing information on individual persons, households or business entities (microdata). The files are created to allow the general public to get familiar with statistical microdata files. The files are prepared in such a way that individual entities cannot be identified.

[2] Scientific-use micorodata files = Microdata Under Contract files (MUC files) meaning de-identified confidential data to which methods of statistical disclosure control (SDC) have been applied (SDC = local suppression e.g. aggregation into age groups, adding noise, synthetic data).

[3] = Research rooms files (RR files) meaning de-identified confidential data to which no SDS has been applied

At the mission the following issues will be addressed:

- *Establishment of a data catalogue including principles and requirements for supporting IT-tools*
- *Documentation incorporated in the data catalogue*
- *How to integrate the data catalogue and a future data library*

2. Expected output of the activity

- *Activity report*
- *Pilots for data catalogue performed and a challenges identified*
- *Policy for a data catalogue outlined*
- *Structure and content of data catalogue outlined*
- *Recommendations on possible steps for building an integrated data library, data catalogues and documentation.*

3. Participants

FOR ALL SESSIONS:

International Relations and Statistical Coordination Department (ICBS) *– Day 1-3*

- Ms. Sigalit Mazeh, BC Component Leader, Director, International Relations and Statistical Coordination Department, sigalit@cbs.gov.il;
- Mr. Adnan Mansur, Head of the Research Service Unit established 01 October 2017), adnanm@cbs.gov.il
- Ms. Anat Katz-Avram, Responsible for research rooms, anatk@cbs.gov.il
- Ms. Amira Abu Rmele, New staff member of the Research Service Unit, amiraa@cbs.gov.il
- Mr. Uriel Naiman, Student, Urieln@cbs.gov.il
- Ms. Areen Aubed Elrachman, Student, areena@cbs.gov.il

Twinning Staff *Day 1-3*

- Ms. Charlotte Nielsen, Resident Twining Adviser cln@dst.dk; CharlotteN@cbs.gov.il
- Ms. Batia Attali, RTA counterpart, International Relations and Statistical Coordination Department; batia@cbs.gov.il
- Ms. Tamar Rand, Resident Twining Adviser Assistant; TamarRa@cbs.gov.il

FOR SELECTED SESSIONS:

International Relations and Statistical Coordination Department (ICBS) *– Day 1-3*

- Ms. Yotal Weiss, Manager of statistical documentation by SIMS, Yotalp@cbs.gov.il

IT and security experts (ICBS) *- Day 1*

- Mr. Itzik Goldstein, Deputy Director of IT, itzikg@cbs.gov.il
- Mr. Genady Etin, Head of the IT Sector, genadye@cbs.gov.il
- Mr. David Gordon gordon@cbs.gov.il

Demography and Census Department *- Day 1*

- Ms. Olivia Blum, Director, blum@cbs.gov.il
- Ms. Liat Rahavi-Italiano, Head of Unit, Population Estimates Sector liatr@cbs.gov.il

Day 1 and Pilot number 1 on day 2

- Ms. Yifat Abuchzira, Population Estimates Sector yifats@cbs.gov.il

Business-Economic Statistics *- Day 1*

- Ms. Tali Nogrian, talino@cbs.gov.il

***Micro-Economic Statistics** - Day 1*

- Ms. Nardit Stein-Kapach, nerdits@cbs.gov.il
- Ms. Larisa Fleishman, larisaf@cbs.gov.il
- Ms. Lior Dopaz, liord@cbs.gov.il
- Ms. Liron Sivan Sherman, lirons@cbs.gov.il

***The forum for research services** - Day 1*

- Ms. Orly Furman, Micro Economic, orlyf@cbs.gov.il
- Ms. Ronit Nissimbaom, Business Economic, ronitn@cbs.gov.il
- Ms. Hadas Yaffe, Census Data, hadasy@cbs.gov.il
- Ms. Yafit Alfandari, yafita@cbs.gov.il
- Ms. Gilat Galimidi, Infrastructure Economic, gilatg@cbs.gov.il
- Ms. Naama Rotem, Health Statistics, naama@cbs.gov.il
- Ms. Nir Fogel, Education and Teaching Staff, nirf@cbs.gov.il
- Mr. Mark Feldman, Labor Force survey, Micro-Economic, feldman@cbs.gov.il

***Representatives from Data Confidentiality Committee and the Data Transfer Committee** - Day 1*

- Mr. Ahmad Hleihel, Deputy Senior Director, Demography and Census Department and Chairman of the data confidentiality Committee since 2011, ahmad@cbs.gov.il
- Ms. Tali Tal, Director of Infrastructure Economic Statistics Department, tali@cbs.gov.il

***The Government Statistician office (ICBS)** - Day 1*

- Mr. Brian Negin, ICBS legal Adviser, briann@cbs.gov.il

Experts from Statistics Denmark

- Expert 1: Charlotte Leolnar Reif, Senior Adviser, Research Services, Statistics Denmark. Contact information: clr@dst.dk
- Expert 2: Karin Holst Duer, Chief Adviser and Project manager for strategic projects in Statistics Denmark, The Management Office, Statistics Denmark. Contact information:khd@dst.dk

4. Current situation in ICBS

Currently 60 active research projects are hosted by the Israeli Central Bureau of Statistics (ICBS). In addition there are about 15 additional research projects in different stages of approval. In total about 100 researchers have access to the current 60 projects distributed among 25 different research Institutions.

Data and data access:

Presently ICBS provides micro-data files for researchers solely in research rooms managed by ICBS. The research rooms are equipped with standalone computers.

Currently de-identified micro-data files are delivered as:

- ***PUF files (Public use microdata files)*** – defined as sets of records containing information on individual persons, households or business entities (microdata). The files are created to allow the general public to get familiar with statistical microdata files. The files are prepared in such a way that individual entities cannot be identified;
- ***MUC files (Microdata Under Contract/scientific-use files)*** – defined as anonymized confidential microdata for research purposes that have been subject to statistical disclosure by aggregation, in order to minimize risk of identification to a low level.
- ***'Research room' files (secure-use files)*** – defined as confidential microdata for research purposes to which only limited methods of statistical disclosure control have been applied.

In order to create the files mentioned above the subject units have databases with identified microdata. Policy, format, structure, content and naming is, however, not coordinated or harmonized between units and sometimes not even within a unit and between years.

For each project *one* customized research dataset is created according to the *need to know principle.* The research dataset integrates multiple datasets delivered by multiple units and may also include external data (from outside the ICBS - data from other administrative authorities and the researchers own data e.g. own surveys).

Current workflow for creation of research datasets:

At the initial request for a research project, one staff member from an ICBS subject matter unit is appointed to be the person in charge for that specific research project. It is the responsibility of the appointed ICBS staff member to coordinate, collect and prepare data as well as collect the associated documentation.

The organization and workflows required to create a research dataset both internally within the ICBS as well as for the researcher are rather complex because they involve several departments, units, committees and managers from the ICBS. Some of the reasons for this are that, presently, no data catalogue is available for internal or external use and no data repository exists in the ICBS today. Thus in order for the ICBS coordinator to get an overview of what datasets are available from different units the coordinator often needs to call and set up meetings with multiple units.. Furthermore, the current documentation of datasets, variable, codelists and classification is not complete and integrated and does not always meet the researcher’s needs for details. An additional obstacle for a more lean production process is that ICBS has no policy, coordination and harmonization of data across domains. This means that the process time from the moment request is submitted until the access is granted may take up to one year.

Researchers need for a data catalogue and documentation:

In December 2016 and January 2017 the ICBS held meetings with 80 current and potential users of the research services in order to get a better understanding of their needs and expectations for future services. One of the main complaints regarding the current services brought up by the research community was the lack of an overview of available data to be used in research projects. This statement was supported by a questionnaire where researchers were asked to give priority to a number of planned initiatives in the ICBS. The availability of a data catalogue was rated as the second most important initiative only slightly surpassed by the possibility to know the status of progress of their project. Another important point that was raised by several researchers was harmonizing names of variables and codebooks across domains and years. Currently a variable can occasionally be named differently in different years, and it is not always clear for the researcher if the name change is due to change of definition or just a change of variable name without any changes of the content.

Catalogue and documentation available for researchers:

In order to get an overview of available data from ICBS, currently, the researcher's option is to look at the ICBS website. However, in most cases the information provided on the webpage is not sufficient for the researcher's needs. Therefore, the researcher needs to consult with the staff members of the subject units that in most cases will customize catalogues and documentation on a case by case basis for each individual research project.

Below a short summary of each platform is listed and examples of the documentation provided in each platform is attached to this ToR as annexes.

Publications and Products Catalogue http://www.cbs.gov.il/reader/publications/bysubject_e_new.htm#6

Publication structure by topics (subject areas)

1. Agriculture
2. Balance of Payments, Foreign Trade and Energy
3. Construction
4. Crime / Public Order
5. Culture, Entertainment and Sport
6. Education
7. Environment
8. Finance and Insurance
9. Geophysical Characteristics
10. Government and Local Authorities
11. Health
12. Immigration
13. Internal Migration
14. Labour and Wages
15. Living Conditions and Welfare
16. Manufacturing and Commerce
17. National Accounts
18. Population
19. Prices
20. Research and Development
21. Social Statistics
22. Tourism and Hotel Services
23. Transport and Communications
24. Vital Statistics

Statistical Abstract of Israel http://www.cbs.gov.il/reader/shnatonenew_site.htm
Each year the ICBS publishes Statistical Abstracts of Israel. The abstract consist of 27 subject oriented chapters with an introduction to each chapter that includes explanatory notes, definitions and sources. An example can be found in Annex A.

Dictionary of Terminology http://www.cbs.gov.il/reader/Milon/Milon_E.html?OnlyFinal=1
The ICBS provides a dictionary of terminology that includes definitions of codes. An example can be found in Annex B

Classification http://www.cbs.gov.il/classifications.htm
The classification currently holds the flowing classifications:

- Standard classification of occupations 2011
- Standard industrial classification of all economic activities (updated edition) 2011
- The conversion key from the 1994 standard classification of occupations to the ISCO 88 classification
- The standard classification of occupations 1994

Customized documentation not publicly available
Currently, customized documentation is created for each delivery of PUF files, MUC files and RR files on a case by case basis. The degree of attention to documentation varies a lot between units and there is no shared repository for documentation except for the dictionary that is only voluntarily used. However, as part of the last Twinning project the ISOPED framework was implemented for educational statistics in the ICBS.

List of Israel Social Science Data Center (ISDC) at Hebrew University of Jerusalem:
The Israel Social Science Data Center (ISDC) provides a concise list of ISDC holding of research datasets (PUF and MUC) on their webpage. This page enables the user to promptly browse the list and locate a dataset. The list is sorted by major keywords. By clicking on a keyword, the researcher can see the relevant datasets and click on a dataset number to read it's abstract and browse its variables through the [record] icon, Please consult: http://isdc.huji.ac.il/ehold4.shtml

On-going development projects:
The new webpage
The ICBS is working on finalizing a complete renewal of their webpage www.cbs.gov.il. The renewal will include a new hierarchical structure of the topics (subject areas) in order to improve usability, user orientation and clarity on the new website (Annex C). In the future all publication, dissemination of data, and documentation will use the new structure of topics.

SIMS (Pilot on educational statistics performed)
The ICBS is in the process of systematizing metadata for their statistical products. The current status is that the ICBS has completed a pilot on education statistics. The pilot used Eurostat's Single Integrated Metadata Structure (SIMS) standard aligned with the quality dimensions of the CoP as a basis for statistical documentation of the statistical products. ICBS applied the Software tool Magic in order to manage the documentation of the statistical output. Most recently the system has been updated to version 2 and front-page fields[4] used by Statistics Denmark has been discussed and might be introduced in the next step.

Metadata are managed by the Metadata Steering Committee and the Metadata Implementation group. Until recently the daily management and development of the project has been carried by a specialized small unit of

[4] Statistical presentation, statistical processing, relevance, accuracy and reliability, comparability, accessibility and clarity

two persons headed by the Director of Economic Research in the National Statistician's Office. Since 01 October 2017 the responsibility for metadata has been transferred to International Relations and Statistical Coordination Department, headed by Director Sigalit Mazeh who is also the Director for research services at the ICBS.

Establishment of a Research Service Unit:
In order for ICBS to provide a more professional and user-friendly service to the researchers a Research Service Unit was created and launched on 01 October 2017. The new unit is managed by Mr. Adnan Mansur under the International Relations and Statistical Coordination Department. In addition to Mr. Adnan Mansur, the unit currently has one additional staff member, Ms. Amira Abu Rmele and a student, Uriel Naiman. However, initiatives to expand the unit are in progress.

5. Current situation at Statistics Denmark

Data catalogue
In Denmark multiple datasets dates back to the 1980’s, which make them very relevant for e.g. longitudinal research projects. The content of the registers also covers many topics e.g. labour market, sociology, health and business. In order to assist the researchers in selecting relevant data for their research project, the research service unit at Statistics Denmark maintains a data catalogue that is available internally but is also made available for the researchers on the internet. The catalogue is updated on a nightly basis through automated readings of the content of a research data library managed by the research service unit. The catalogue contains all data produced by Statistics Denmark as well as administrative data from other public agencies. Inclusion of datasets from other national agencies is ensured by bilateral agreements between the national agencies and Statistics Denmark.

In summary the catalogue contains 273 unique file names, ~5000 datasets and more than >12.000 unique variables. The catalogue is maintained as an interactive catalogue with hyperlinks to the documentation at both register and variable level.

Metadata for researchers
In Denmark the production of metadata and documentation is the responsibility of the subject units. The publication strategy at Statistics Denmark states that no data can be published form Statistics Denmark before documentation of the statistical product (dataset) and variables have been accurately documented. The documentation is made available for internal users (intranet) and external users (internet) at the same time as data are published. The policy applies to all publication channels used in Denmark (Statbank, Notes, Press realises, NEWS (NYT’er) etc).

Documentation of datasets - All statistical products published by DST are described in Documentation of statistics that can found on the internet while working on the research servers and for internal DST employees on the documentation system Collectica. The documentation is based on the international standard SIMS. Altogether DST’s production includes more than 250 statistical products and is documented in both Danish and English.

Documentation of variables - All variables in DST’s statistics are documented in the documentation system TIMES (currently in the process of being transferred to Collectica). TIMES consists of a hierarchically constructed common database with associated variable's description. TIMES documentation is available only in Danish.

High Quality Documentation - In addition a special documentation format particularly aimed for researchers was initiated in 2008, the so called High Quality Documentation started due to demand of the research community.

The new element in the High Quality Documentation is the inclusion of the historical dimension in the documentation for each register and variable. The ***historical dimension*** describes possible data rupture and validity periods and the population is specified for each variable. Where applicable a graph and a table are attached in order to illustrate the historical development. The High Quality Documentation is produced in corporation between researchers, subject units and the Research Service unit. What is also unique about this documentation is that it includes a ***review process*** by a prominent researcher within the subject area. Since

2008 the cost of the High Quality Documentation has been covered by a yearly grant from the Danish Agency for Science and Higher Education.

In order to assist the High Quality Documentation a guideline for the process and content was created for the subject unit by the research service unit.

As for the regular documentation the TIMES system is used for the High quality documentation, however, a few additional fields applied in the template in order to include the extra elements.

Codelist/format library - Codelists and a format library are available on the research servers. A practical guideline on how to use the format library is provided for researchers. Currently, the codelist and classification are in the process of being integrated in Collectica where codelists and classifications are documented in CSV and DDI[5] format.

Feedback from the research community:

The external review process for High quality documentation - The external quality assurance for High quality documentation consists of evaluating all variables within a single area by two external reviewers. These reviewers are persons who have special insight into the area concerned. The reviewers go carefully over the documentation and make suggestions for corrections, clarification needed, etc. In cases where there are annexes, in the form of documents produced, attached to a variable, the external evaluators will also comment on these. In addition, external reviewers may collect general remarks to a registry area in a single document, which is sent to Research Services together with external assessments.

KOR - The Coordinating Body for Registry Research (KOR) is an advisory body under the Ministry of Education and Research, with reference to the Board of Research and Innovation.

The goal of KOR is to stimulate and strengthen Danish registry research. KOR contributes to creating greater coherence and coordination around both Danish and internationally related research activities concerning registers, databases and survey data. KOR consists of active researchers from research and university environments. KOR administers a fiscal license for register research, which is among others used for the operation of the research service at Statistics Denmark. Agreement of datasets and variables to be High Quality documented is selected on a yearly basis by KOR based on input from the research service unit.

The Research Forum – an internet forum for researchers that was hosted by the research service unit with the aim of sharing knowledge and raise questions concerning datasets, variables etc. The research service hosted this forum for a number of years but closed it down in 2011 since it was hardly used by the research community, not even after the research unit in cooperation with some selected researchers tried to boost the activity by posting questions and answers for a period of time. The researchers stated that the idea of knowledge sharing was brilliant, but their time was too sparse and it took too long to answer other researchers. Thus they preferred to get advice directly from the research service unit.

The Research Committee - The Research Committee is an advisory committee dealing with general questions in relation to Statistics Denmark and society research. The Committee also assists strengthening the contact to the research community and thereby contributes to a better utilization of basic data in social research. The committee meets on a regularly basis two times per year and consist of 12 external members from the

[5] The Data Documentation Initiative. DDI is an international standard for describing surveys, questionnaires, statistical data files, and social science study-level information. This information is described as metadata by the standard. DDI began in 1995 and brought together data professionals from around the world to develop the standard. The DDI specification, written in XML, provides a format for content, exchange, and preservation of questionnaire and data file information. DDI fills a need related to the challenge of storing and distributing social science metadata, creating an international standard for the design of codebooks. The freely available international DDI standard describes data that result from observational methods in the social, behavioral, economic, and health sciences.

research community and three internal members from Statistics Denmark. Workgroups may be set up in connection with the committee.

The members are appointed by the National Statistician based on recommendation of the research service unit. Members are selected in a way that enables as wide a spread as possible between institutions and research areas. The main topics are economics, labour market research, social medicine, demography, sociology, social research and statistical methods. The committee's committee and member circle are reviewed at least every three years.

The committee addresses the following issues:

- Advice on all statistical areas in their work plan.
- Advise Statistics Denmark on their confidentiality principles in relation to research projects
- Advice on individual projects that Statistics Denmark should focus on
- Advice on statistical methodological issues.
- Advise in dissemination (databases, metadata, publications etc.)

Satisfaction surveys – At least once a year all active researchers receive an online satisfaction survey in order to constantly improve the service provided by Statistics Denmark. One of the questions regards to their satisfaction with the documentation and suggestions on how to improve it.

6. Vision for building and integrating data library, data catalogue and documentation

The vision for future services for researchers provided by the ICBS is (i) to provide microdata and services in a secure, professional, user-friendly and competent manner meeting international standards and best practices and (ii) to increase effectiveness in order to reduce the time spent from when a request is placed and until access to data can be obtained.

The vision is to build a standardized and fully integrated data catalogue, data library, and documentation in a stepwise approach. The first step being outlining a policy, the structure and content of a data catalogue that are standardized and harmonized across domains and years that can be used by the researcher (i) to select datasets and variables suitable for testing their research hypothesis, (ii) to be used when working with their data, and (iii) to be used in writing up their results (before/under/after). The development need to be done in close cooperation between the subject units and the IT department.

During the development phase it is particularly important to emphasize the long term benefits for the statistical production units. In particular to demonstrate that when this new approach is fully developed the subject units only need to produce one research dataset for each statistic and to document the statistics once and NOT every time there is a request for a research project, since the data library and documentation will be built in such a way that research data and documentation can be used by the research service unit to build customized datasets and documentation for each project.

The data library should be built with no direct identifiers and standardized and harmonized across domains and years. The end goal is that the work of providing standardized datasets for research will be a natural and integrated part of the subject units obligation for producing statistics and that the process for producing research data sets is done according to the General Statistical Business Process Model (GSBPM).