Statistical Microdata – Production of Safe Datasets, Transparent Presentation of Contents and Advanced Services for Users through Metadata Authorization System

M. Poljičak*, H. Stančić**
* Central Bureau of Statistics/Information System Design and Development Department, Zagreb, Croatia
** Faculty of Humanities and Social Sciences/Department of Information and Communication Sciences, Zagreb, Croatia
*poljicakm@dzs.com
**hstancic@ffzg.hr

In the European Statistical System (ESS) there are more and more generic information systems (ISs) being developed and implemented for production and dissemination of official statistics (OS) macrodata datasets. These generic systems are based on metadata repositories providing, in return, a whole set of additional functionalities and services for users of the systems. Microdata dissemination, being the youngest and least developed discipline in official statistics dataset dissemination, currently still lacks appropriate organization of business procedures and compatible generic tools for the purpose of integration of all the ESS' subsystems into one central Data Service (DS) providing microdata datasets for international users and enabling preservation of microdata datasets. The authors analyze current situation regarding statistical microdata sets acquisition and preservation in digital archives (DAs), provide valuable considerations and propose recommendations to take into account when developing repositories and systems for presentation of available microdata datasets to users. Along with that, the authors introduce the way of usage of appropriate Statistical Disclosure Control (SDC) methods for production of safe microdata outputs. Finally, the authors recommend the development of a central administrative metadata repository that could be used for users' authorization, and that could have additional services for users based on available metadata about users in the system.

I. INTRODUCTION

Protecting the confidentiality of data in statistical surveys is one of the *Fundamental Principles of Official Statistics (OS).* OS should care about secure handling, preservation and dissemination of data. Setting authorized access levels as well as not publishing any data in any way that could reveal information, protected by legal or ethical restrictions, is necessary.

When it comes to data preservation in Official Data Archives (ODAs) different situations are present, at the moment, in different European countries. Some countries have a considerable level of cooperation between National Statistical Institutes (NSIs) and ODAs, and others still haven't developed satisfying and proper levels of cooperation and legal, administrative and technical solutions to tackle these needs.

It is of great importance to maintain the trust of the public in OS, and this is accomplished by setting extra efforts in regulating ways of accessing data, with proper security measures put in place when access is granted.

The problem of setting security measures and protocols, as well as secured ways of access to data, is complex especially if one takes into account a wide range of statistical data outputs available to users. These data outputs have different confidentiality levels and disclosure risks.

Statistical Disclosure Control (SDC) methods take into account legal and administrative boundaries to assure dissemination strategy aligned to disclosure risk and management of the same disclosure risk. Usage of appropriate IT tools and technical solutions, aligned with needs of legal compliance of procedures involved in dissemination of data, is still in the development phase in some areas. In these procedures a lot of manual work is done, involving time of skilled professionals in performing SDC methods, i.e. checking data output for statistical confidentiality. The area of microdata needs a better level of automatization of procedures in order to achieve a faster and simpler channel for production of safe data outputs for research community use.

Currently, new services, aligned to the new orientation of governments across the globe for opening access to different available government datasets, are calling for the new establishment of OSs' dissemination strategies. Organizations responsible for the development of OS in countries should invest efforts in establishing better ways of communication between NSIs, ODAs and research community (universities and research institutes in countries and abroad). These organizations should receive the support of political and other influential bodies in the country, as well as all necessary resources, to implement the procedures and prepare all protocols needed to raise the level of availability of microdata. Raised availability of microdata, could be useful and contribute to better de-

cision making and policies setting for the society. In effect they could bring new and advanced environment development and decision making concerning industry, economy and all other important sectors for empowering the society. Some of the examples of interested users of OS data are economists, demographers and social science researchers, as they could benefit from using vast resources created and preserved in conducting statistical surveys. All these users can find new and advanced secondary uses of extensive and expensive material already collected [12]. However, if data are released to these users, it must be under a proper disclosure control level in place for confidential statistical data.

The broad classification of statistical outputs usually distinguishes between two categories – *macrodata* and *microdata*.

Macrodata are aggregated statistical data (sums, averages, counting units), and these data are usually publicly available. However, some macrodata sets could be restrained from dissemination to public because the sample is too revealing and therefore public could infer conclusions about individual records in the macrodata dataset.

On the other hand, *microdata* are records collected as input data for the surveys. These records are confidential, subject to law restrictions and ethical standards set in place. However, under special conditions these data could be released, with more or less content (anonymized datasets), to special groups of users. Following current regulations, the users of microdata datasets could be any person belonging to the group of university researchers or members of other research institutions in the country, or even outside of the country, if they ask for accreditation. The procedure and the level of access rights are different in each European country, as are the ways of access. In the year 2013 a new regulation by the European Commission EC 557/2013 is set in place. By that regulation the process of application for statistical microdata access is made transparent for research community. "This Regulation establishes the conditions under which access to confidential data transmitted to the Commission (Eurostat) may be granted for enabling statistical analyses for scientific purposes, and the rules of cooperation between the Commission (Eurostat) and national statistical authorities in order to facilitate such access." [6]

A lot of work is still needed for the establishment of a successful Central Remote Access Data Center (RADC) for European microdata, providing the researchers with a simple and comfortable access to microdata and transparent presentation of data available. The service is currently being developed by the Eurostat and it should provide researchers with a central RA microdata access point in Europe.

In this paper the authors draw conclusions about several issues relating to establishing RADCs.

(1) How automated SDC procedures should be developed through RADC and aligned to the disclosure risk management for microdata datasets,

(2) How available resources should be transparently shown to the users, organized and made easily searchable,

(3) How researchers' authorization levels should be managed in a way to provide users with additional services aligned to their interests and needs concerning statistical microdata.

II. Related work

In providing public, and especially research community, with valuable statistical information NSIs need to set the balance between the level of the statistical outputs released and levels of protection of confidentiality of the individual data of respondents. In order to set an appropriate SDC method in place, i.e. to choose and implement appropriate software solutions, providers of official statistics microdata (as well as aggregated data outputs) should have in mind who their users are, as well as the possible uses of the released data. Also they need to consider possible further dissemination of the released data. For dissemination of the released data a proper ways of presenting available data should be defined. Also, users of the RADC, i.e. authorized researchers, should have an easy way of authentication and authorization procedures in order to manage data retrieval and to be able to get all other advanced services needed from the system.

A. Statistical Disclosure Control

In [1] "To protect the privacy of respondents a statistical office should prevent the disclosure of sensitive information on individual respondents by an intruder. As disclosure of sensitive information is possible only after an individual has been identified, a statistical office usually tries to prevent the identification of individual respondents."

As in [3] "Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL) seeks to protect statistical data in such a way that they can be released without giving away confidential information that can be linked to specific individuals or entities." Also in [3] "a disclosure occurs when a person or organization recognizes or learns something that they did not know already about another person or organization via released data." And also in [3] "statistical disclosure control techniques can be defined as the set of methods to reduce the risk of disclosing information about individuals, businesses or other organizations." SDC methods minimize the risk of disclosure to an acceptable level while releasing as much information as possible. There are two types of disclosure risks as in [3], *identity disclosure* and *attribute disclosure*. Identity disclosure occurs with the association of a respondent's identity with the disseminated data record containing confidential information. *Attribute disclosure* occurs with the association of either attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent.

As mentioned in [3], microdata protection methods used for generating safe datasets can be divided into two categories: *masking methods* and *synthetic data generating methods*. *Masking methods* mask the original data to protect the dataset from the risk of disclosure. *Synthetic data methods* preserve some statistical properties of the original data, but on the other hand, are difficult to implement.

There are two types of SDC masking methods [3] – *perturbative and non-perturbative methods.* Perturbative methods falsify the data before publication by introducing an element of error purposefully for confidentiality reasons. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset. Such confusion is beneficial for preserving statistical confidentiality. The used perturbation method should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. On the other hand, non-perturbative methods reduce the amount of information released by suppression or aggregation of data. Global recoding, local suppression and sampling are examples of non-perturbative methods." In that sense "*global recoding* method is a method where several categories of a variable are combined into single one, and *local suppression* method is a method where a value of a variable in a record is replaced by *missing*." [1]

In [3] it is showed that "a microdata set can be viewed as a file with *n* records, where each record contains *m* variables on an individual respondent. The variables can be classified in four categories which are not necessarily disjoint:

(1) *Identifiers.* These are variables that unambiguously identify the respondent (in Croatia, the example is OIB – Personal Identification Number, author's comment).

(2) *Quasi-identifiers or key variables.* These are variables which identify the respondent with some degree of ambiguity, but they can provide unambiguous identification of the individual record. Examples are name, address, gender, telephone number, etc.

(3) *Confidential outcome variables.* These are variables that contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

(4) *Non-confidential outcome variables.* These are variables which do not fall in any of the categories above."

Datasets anonymization could be done by avoiding dissemination of identifiers and quasi-identifiers since some of the general conclusions about the type of data are considered highly disclosing. Besides that, in anonymization procedures usually "categories of identifying variables with too significant identifying power are commonly aggregated into a single category. (...) Geographical information is a strongly identifying variable. (...) Release of date of birth is highly discouraged." [3] Also, when dealing with business data it is very easy to find the data belonging to a largest national company in some sectors, i.e. income or turnover of the company in telecommunications or national oil and gas company, so this data should be specially protected, especially in smaller countries where there is not a lot of suppliers of some services or goods in some sectors. "Examples of identifying variables are Age, Sex, Domicile and Occupation. Although each identifying variable is generally not sufficient to identify an individual when considered separately, a combination of values of identifying variables might be sufficient. When a combination of values of identifying variables is unique, i.e. occurs only once in population, then an intruder might identify the corresponding individual. (...) For preventing disclosure it is better to avoid the occurrence of combinations of values in the microdata set that are rare in the population, instead of trying to avoid only the population-uniques in the microdata set." [1]

The rules for handling disclosure risk take into account settings of the threshold, i.e. the minimum number of occurrence of the value of variable, considered to be the minimum for allowing the variable to be released in the output microdata dataset. "In case the frequency of a particular combination is at least the prescribed threshold value then this combination is considered safe. Otherwise the combination is considered unsafe and disclosure limitation measures should be applied." [1]

Access and overall dissemination of identifiers and quasi-identifiers should be avoided, especially taking into account possibility of merging these data with other available databases and greater risk levels concerned with possible disclosure of information in that scenario. Later in this research paper we will suggest providing free access to non-confidential outcome variables, with possibility of protecting confidential variables and/or confidential sets of variables. Restricted ways of access to confidential variables using SDC methods will be analyzed.

B. Presentation of data available in RADC

The second point of interest of this paper is the presentation of available content in an easy to search and explore ways, i.e. using navigation tools in RADC with possibility of learning about available datasets.

Developing a central RADC in Europe will not be an easy task. There are many countries with their NSIs involved in collecting, processing, preserving, and disseminating statistical datasets. Many of NSIs are either in the process of implementing standards for data collecting, documenting and preserving of the collected and processed data, or they have already done it. This should ease the situation in the future and enable a better connectivity between resources from different countries. Statistical surveys are also becoming standardized in the methodology for collecting, processing and disseminating data and in data structures.

C. Authorization system using metadata repository

The third point of interest of this paper is the ways of enabling advanced services for users of RA to microdata, as well as providing authorization mechanisms according to the authorization levels of users of the system and according to the authorization level needed to perform specialized operation on a dataset. It is obvious that in this DC, enabling access to confidential data, only authorized users should have access to data. The access should be granted according to all legal, technical and other measures taken into account and set in place for the proper level of data protection. This can be achieved by using metadata repository of users. Such repository should be placed in the core of the system responsible for documenting users, their rights and prescribed services available to a specific user.

III. MICRODATA SDC

A data dissemination strategy offers many different statistical outputs covering a range of different topics for many types of users. Different outputs require different approaches to SDC and different mixture of tools for handling it.

The microdata require a high level of disclosure control in order to protect confidentiality of available datasets. Usually anonymized datasets are available to users. Users can only perform statistical analysis on data – they cannot see the real data. The produced outputs are deposited in the system's repository. These outputs are first checked by NSI staff for statistical disclosure, and only if approved, the secondary data retrieved by using microdata can be released and used by the researcher.

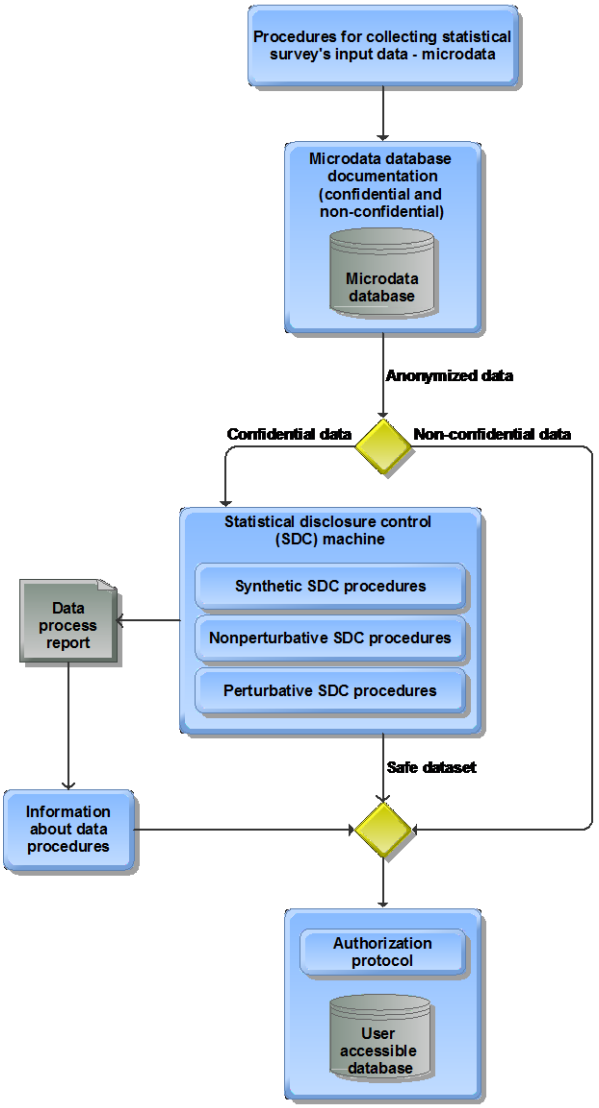
From the new Regulation introduced by the European Commission – EC 557/2013 – it can be concluded that the process of granting access to statistical microdata is becoming more transparent and procedures involved are more organized and centralized. In [6] "The Commission (Eurostat) shall publish on its website: (a) guidelines for the assessment of research entities, research proposals and access facilities; (b) the list of recognized research entities; (c) the list of accredited access facilities: (d) the list of datasets for research use with relevant documentation and the available modes of access."

However, for the NSI's staff, SDC procedures are still very time-consuming and burdening.

Therefore, we suggest building of a system for automatic creation of safe statistical outputs in the form of appropriate level of anonymization. Also, we suggest that variables and/or sets of variables considered to be revealing before being released through RADC, could be handled by development and implementation of software solution for enabling extraction of safe portions of datasets, as well as providing transparent documentation about applied SDC methods. Those datasets would then be made available to users in the form of a safe dataset with collection of non-confidential attributes and SDC protected confidential attributes which can then be used in a more user-friendly way, without creating barriers between users and the real datasets.

This system could be used for microdata dissemination strategy implementation for the research community. Accomplishing that, it is only a question of the extent of usability of the received datasets for research which are in accordance with the need for assuring the highest level of usability of data while at the same time protecting confidentiality of datasets.

However, there are, at the moment, software solutions for SDC. One example is µArgus, developed at Statistics Netherlands in 1995, using various SDC methods. The application is designed to work through several phases for ensuring production of safe data output. It works interactively with the user of the system. The metadata about data should be introduced to the system at the beginning of the process, and at the end the system produces a data process report for the data entering µArgus production of safe dataset.



Picture 1. Production of safe microdata outputs with documentation on applied SDC methods

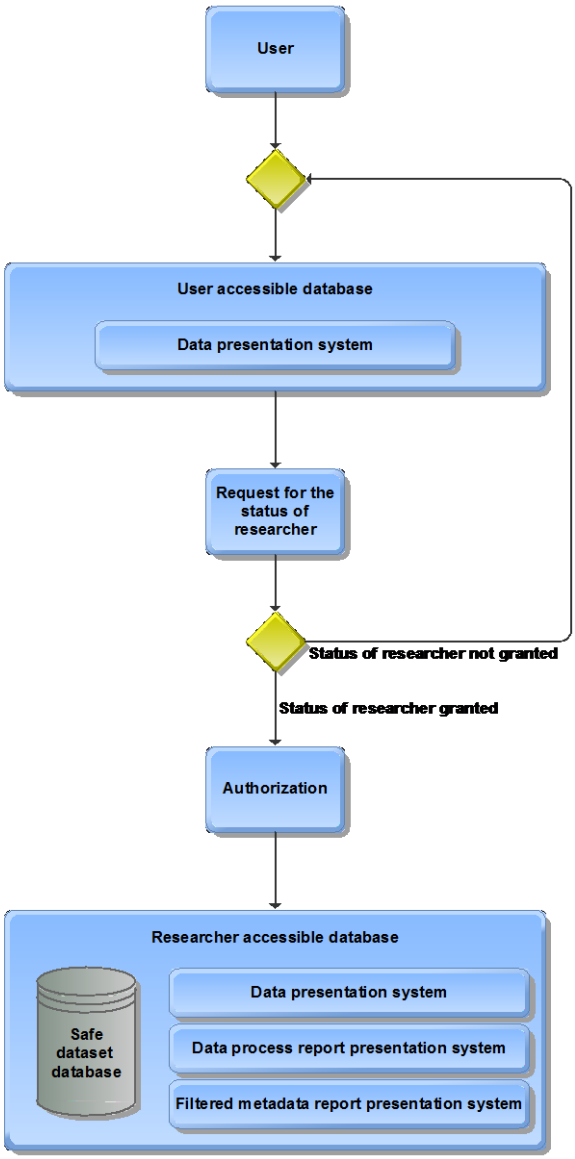
IV. TRANSPARENT PRESENTATION OF DATA IN THE USER ACCESS SYSTEM

When procedures for obtaining data collection and the documentation of the collected data in a Digital Archive (DA) become established for a statistical survey, then the same statistical survey data documentation should become available to the researchers through the DA interface that enables easy navigation and retrieval.

It is customary to use metadata repositories and semantics methodology for presentation of available resources in an interactive and easily retrievable way.

Metadata about surveys and datasets can help users learn about the available datasets and the survey methodology, and also help them to explore vast materials they are not familiar with.

The possible ways of showing the available contents are through interactive controls, like navigation trees, which should be prepared as a multilingual environment, thus enabling access and usage for possible users coming from different European countries. This goal is achievable if the system is generically built with resource-based approach to documenting all the data and metadata in a translatable repository of metadata. This repository can later be used to manage building multilingual controls for the user interface according to the available categories of data.



Picture 2. User access system

"Documentation is an essential part of any dissemination strategy both for auditing from external authorities and transparency towards users. The former may include descriptions of legal and administrative steps for a risk management policy together with the technical solutions applied. The latter is essential for a user to understand what has been changed or limited in the data because of confidentiality constraints. If a data perturbation method has been applied then, for transparency reasons, this should be clearly stated. (...) If a data reduction method has been applied with some local suppression (...) The released microdata should be obviously accompanied with all necessary metadata and information on methodologies used at various stages of the survey process (sampling, imputation, validation, etc.) together with information on magnitude of sampling errors, estimation domains, etc." [3]

V. USER AUTHORIZATION THROUGH METADATA REPOSITORY

By using a metadata repository for the purpose of building authorization system for diverse users, it is possible to develop all kinds of additional services for users, like subscribing to information about new datasets available, informing users about expiration of subscription to the data in advance, etc.

A metadata repository should hold the data about a user, surveys available to him/her to explore or use, and data released for the user exploration or additional analysis. This means that the repository will, in return, have a history of the user rights to datasets – preserved through time, and also it should be possible to record usage of the data to the user profile. This information could possibly be used in cases of confidentiality breech as well as for informing the user about some important information, or information concerning the deposit of new data and similar functions. In this way system's resources could be personalized in accordance with the users' interests and needs.

In spite of being generic and very versatile regarding available services for the user, this system should be easy to use. It should enable easy creation of new items in the database, as well as inheritance of all available modules. It should also be practical for the purposes of reuse and updating.

VI. CONCLUSION

The aim of this paper is to answer three questions important to have in mind when building either a central international RADC for statistical microdata access for researchers, or a national system for the same purpose in a national context.

Concerning the first question stated in the Introduction part of the paper, it can be concluded that SDC methods, aligned with the dissemination strategy for the data output, should be implemented through an automated system. Such a system should provide researchers with a satisfying level of data usability, and cover disclosure protection according to the level of risk of identity or attribute disclosure. This could be done by the proposed methods of anonymization and suppression of confidential values of variables where considered necessary, as well as masking of all confidential values of variables in the microdata output. Along with mentioned methods global recoding of confidential sets of variables in an automated system should be used in order to get the safe microdata dataset. Sophisticated methods, along with enabling a better level of data usability, should use perturbative methods for enabling non-disclosure scenario, because the data would be

"false", thus eliminating the disclosure risk, but at the same time "true" for obtaining exact aggregated statistical indicators.

The second question yielded a conclusion that the system providing transparent presentation of resources should be developed in a generic way, using metadata repository available in different languages, as well as using semantics methodology for building user interfaces that should enable easy exploration by users, as well as learning about the available datasets in overall transparent presentation of available resources in different languages. This system should certainly provide users with information about SDC methods used with the data before its dissemination.

Answers to the third question are in line with the usage of metadata repository for users' authorization levels implementation. They also enable the development of all kinds of advanced services for the users through personalized profile, developed on the grounds of users' metadata information. This system should be developed using a generic approach incorporating translatable metadata repository in the core of the system.

LIST OF ACRONYMS

DA	Digital Archive
DS	Data Service
ESS	European Statistical System
NSI	National Statistical Institute
ODA	Official Data Archives
OS	Official Statistics
RADC	Remote Access Data Centre
SDC	Statistical Disclosure Control
SDL	Statistical Disclosure Limitation

REFERENCES

[1] A.G. de Waal and L.C.R.J. Willenborg, "Optimal Local Suppression in Microdata", Journal of Official Statistics, 1998, http://www.jos.nu/Articles/article.asp

[2] A.Hunderpool and L.C.R.J. Willenborg, "m- and t- ARGUS: Software for Statistical Disclosure Control", Record linkage Techniques, Chapter 5, 1997, http://www.fcsm.gov/working-papers/hundepool.pdf

[3] A. Hunderpool et al., "Handbook on Statistical Disclosure Control, Version 1.2" ESSNet SDS, 2010, http://neon.vb.cbs.nl/casc/SDC_ Handbook.pdf

[4] B. Schouten and M. Cigrang, "Remote Access Systems for Statistical Analysis of Microdata," Voorburg/Heerlen July 2003, Discussion paper 03004, http://www.cbs.nl/NR/rdonlyres/C8B18053-7B3E-485F-851E-A0DE40350553/0/Discussionpaper03004.pdf

[5] Data without Boundaries, "Deliverable D7.1, Metadata Standards – usage and needs in NSIs and Data Archives," 2012, Work package 7, Standards Development http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d7-1_metadata-standards-usage_report.pdf

[6] European Commission, "Commission Regulation (EU) NO 557/2013 on access to confidential data for scientific purposes," 2013, Official Journal of the European Union, L 164/16, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

[7] European Commission and Eurostat, "Guidelines for the assessment of research entities and research proposals," 2012, http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/documents/guidelines_assessment.pdf

[8] European Commission and Eurostat, "List of recognised research entities," 2012, http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/documents/Recognised_research_entities.pdf.

[9] F. Willis-Nunez, "Users of the statistical metadata system", Statistical metadata (METIS), 2012, http://www1.unece.org/stat/platform/display/metis/3.++Users+of+the+statistical+metadata+system

[10] G. Pongas G. and A. Wronski, "A distributed architecture for statistical data editing, processing and dissemination," Meeting on the Management of Statistical Information Systems (MSIS 2009), Oslo, Norway, 2009, Topic (iii): Architecture, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/35_France-Germany-NL.pdf

[11] N. Ahmad, Koen De Bacher and Y. Yoon, "An OECD perspective on microdata access: Trends, opportunities and challenges," Statistical Journal of the IAOS 26(2009/2010) 57-63, http://donnageczi.com/pdfs/IOS-Press-Journal-Library/support-files/sji.pdf

[12] M. Poljičak and H. Stančić, "Proposing the Model for Croatian Remote Access Safe Centre for Statistical Microdata", in: A. Gilliland et al., Information Governance, Zagreb : Department of information and Communication Sciences, Faculty of Humanities and Social Sciences, 2013, pp. 137-146, http://infoz.ffzg.hr/INFuture/papers/4-02%20Poljicak,%20Stancic,%20Proposing%20the%20Model%20for%20Croatian%20RA%20Safe%20Centre.pdf

[13] R. Grim, P. Heus, T. Mulcahy and J. Ryssevik, "Secure remote access system for an upgraded CESSDA RI, metadata technology, Version Sep 2009" cessda ppp, http://www.cessda.org/project/doc/CESSDA_RI_SRA_FINAL.pdf

[14] R. Silberman, S. Bender and A. Hunderpool, "The need for networks on data access – Data Without Boundaries Project and the Workshop on data access," Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona-Spain, 26-28 October 2011, Topic (vii): Trans-border access to microdata, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/35_France-Germany-NL.pdf

[15] S. Bender, J. Heining, L. Franconi and D. Ichim, "Microdata access: and international perspective," Tarragona-Spain, 26-28 March 2013, BLUE-Enterprise and Trade Statistics, SP1 – Cooperation-Collaborative Project, Small or medium-scale focused research project, http://www.unece.org/fileadmin/DAM/stats/documents/ecen/ces/ge.46/2011/35_France-Germany-NL.pdf

[16] U. Jensen, "Data and metadata extensions of the CESSDA RI (D8.3)," CESSDA PPP Project Deliverables 2010, http://cessda.org/project/doc/D8.3_Data_metdata_enhancement.pdf

[17] United Nations Economic Commission for Europe, "Managing statistical confidentiality & microdata access, Principles and guidelines of good practice," Conference of European Statisticians, New York and Geneva 2007, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

[18] United Nations Economic Commission for Europe, "CMF Part A, Statistical metadata in a corporate context: A guide for managers," Conference of European Statisticians, Geneva 2009, http://www1.unece.org/stat/platform/display/metis/Part+A+-+Statistical+Metadata+in+a+Corporate+Context