FDZ-Arbeitspapier
Nr.5

Joachim Merz
Daniel Vorgrimler
Markus Zwick

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

De facto anonymised microdata file on income tax statistics 1998

2005

FDZ-Arbeitspapier
Nr.5

Joachim Merz
Daniel Vorgrimler
Markus Zwick

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

De facto anonymised microdata file on income tax statistics 1998

2005

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Statistisches Bundesamt

Fachliche Informationen
zu dieser Veröffentlichung:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 06 11 / 75 42 20
Fax: 06 11 / 72 40 00
forschungsdatenzentrum@destatis.de

Erscheinungsfolge: unregelmäßig
Erschienen im Oktober 2005

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 06 11 / 75 42 20
Fax: 06 11 / 72 40 00
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 / 9449 41 47
Fax: 0211 / 9449 40 77
forschungsdatenzentrum@lds.nrw.de

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

De facto anonymised microdata file on income tax statistics 1998

Joachim Merz, Daniel Vorgrimler, Markus Zwick[1]

With the data of the de facto anonymised income tax statistics 1998 (FAST 98), the German official statistics are for the first time publishing microdata from the field of fiscal statistics. The scientific community can use these data to analyse politically-relevant questions on the fiscal and transfer system at their own workplace, subject to the premises of article 16 subsection 6 of the Law on Statistics for Federal Purposes, on the basis of "real" assessment data.

Passing on individual data to the scientific community is only possible in a de facto anonymised form. This form may impair possibilities for scientific analysis possibilities. So that anonymised data can nevertheless be used by the scientific community, anonymisation must meet two equal challenges: It must firstly guarantee adequate protection of the individual items of data, and secondly it must optimally conserve the possibilities for analysis of the anonymised data. In order to achieve the right balance between these two goals, the Statistical Offices have involved potential scientific users in the anonymisation work in a research project.

In the article entitled "De facto anonymised microdata file on income tax statistics 1998", in addition to the anonymisation concept the framework conditions of the project are explained and the analysis possibilities of income tax statistics demonstrated.

1 Introduction

1.1 De facto anonymity of microdata

Microdata are statisticians' raw materials. This personal or factual information on individual respondents, be they individuals, households or enterprises, forms the initial information which is combined in the statistical production process, and which permits a comprehensible portrayal of mass manifestations, for example in the shape of tables. Whilst for a long time into the nineteen sixties as a rule only the Statistical Offices were able to process these mass data, the rapid development of data processing has now made it possible for almost any student to evaluate large volumes of data. Since microdata permit many layers of analysis, the long-standing wish of the scientific community[2] to analyse these data in their original form as individual items of data has grown considerably over time. In 1981, the legislature reacted to the growing need for official individual data, and in article 11 of Law on Statistics for Federal Purpose (Federal Statistics Law - FSL) 1981 created the legal basis to transmit individual data to users outside the Statistical Of-

[1] Prof. Dr. Joachim Merz, University of Lüneburg, Department of Economic and Social Sciences,, Research Institute on Professions (Forschungsinstitut Freie Berufe (FFB)), Dr. Daniel Vorgrimler and Dipl. Volksw. Markus Zwick, Federal Statistical Office

[2] Where this article speaks of "the scientific community" or of "scientists", this refers to scientists working outside the Statistical Offices.

fices. According to this statute, individual data were suitable for transmission if identification of the respondents could be ruled out with absolute certainty. As was shown in practical terms during the ensuing period, the call for absolute anonymity in the individual data led to a situation in which virtually no stocks of individual data were provided to scientists since it was all but impossible to meet this restrictive demand with regard to the data. However, in spite of the strict requirements it was possible in some cases to make data stocks accessible to the scientific community. These however contained so little information that the data were frequently inadequate for scientific research.

As a result of this experience, the Law on Statistics for Federal Purpose (Federal Statistics Law - FSL) was adjusted in the next modification which took place in 1987, comprising the addition of article 16 subsection 6. Accordingly, individual data may be transmitted to the scientific community "if an assignment of the individual data is possible only with an excessive amount of time, expenses and manpower". This "principle of disproportionality" implies that a breach of the anonymity of respondents applies only to beneficial attributions.[3] Hence, the legislature does not require absolute anonymity, instead of which so-called de facto anonymity is considered adequate. Since this only applies to "higher education or other institutions entrusted with tasks of independent scientific research", this regulation is also referred to as a "Privilege of Science".[4]

At the beginning of the nineties, the term "de facto anonymity" was coined in the context of the work carried out by Müller, Blien, Knoche and Wirth[5]. First of all, initial de facto anonymised individual data stocks were transmitted to the scientific community in the shape of the Microcensus, and in the ensuing period with the Sample Survey of Income and Expenditure.

1.2 The Research Data Centres of the Statistical Offices of the Federation and of the Länder

The discussion regarding access to data, and in particular access to individual items of official data, continued after this time. With the memorandum by Professors Richard Hauser, Gerd Wagner and Klaus F. Zimmermann in the *Allgemeines Statistisches Archiv* (Journal of the German Statistical Society)[6], in 1998 the topic took on a new dynamic which ultimately led to the establishment of the Commission for the Improvement of the Informational Infrastructure between the Scientific Community and Statistics. The Commission's final report, published in 2001, contained a wide variety of recommendations to improve the informational infrastructure.[7] In addition to the proposal to establish a long-term Board for Social and Economic Data[8], it was the recommendation to establish Research Data Centres with the data producers in particular which led and is leading to a sustained improvement in access to data for the scientific community.

The establishment of the two Research Data Centres of the Statistical Offices of the Federation and of the Länder in 2001 and 2002 was the direct reaction of official statistics to the Commission's recommendations. The major goal of both Research Data Centres is to facilitate access for the scientific community to the official statistics microdata by expanding existing data access points and by setting up new possibili-

[3] cf. Höhne, J./Sturm, R./Vorgrimler, D.: Konzept zur Schutzwirkung faktischer Anonymisierung, in Wirtschaft und Statistik, 4/2003, p. 287.
[4] On these developments cf. Hans-Jürgen Krupp: Mikroanalysen und amtliche Statistik - gestern, heute morgen, in Merz, J.; Zwick, M. (eds.): MIKAS - Mikroanalysen und amtliche Statistik; Federal Statistical Office, Statistik und Wissenschaft, Vol. 1, 2004
[5] Müller, W.; Blien, U.; Knoche, P. and Wirth, H.: Die faktische Anonymität von Mikrodaten; Vol. 19 of the series of publications entitled Forum der Bundesstatistik, Federal Statistical Office, 1991
[6] Hauser, R.; Wagner, G.; Zimmermann, K.: Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung: Ein Memorandum; Allgemeines Statistical Archiv, Vol. 82, pp. 369-379.
[7] Commission for the Improvement of the Informational Infrastructure between the Scientific Community and Statistics (ed.) "Wege zu einer besseren informationellen Infrastruktur", Baden-Baden 2001.
[8] cf. on this www.RatSWD.de

ties for access. To this end, the Research Data Centres of the Statistical Offices of the Federation and of the Länder, as well as establishing Safe Scientific Workstations in the protected premises of official statistics, have taken a significant step towards controlled remote computing with the further creation of scientific-use files.[9] The 'De facto anonymised microdata file of income tax statistics 1998 (FAST 98)' came into being as a result of the cooperation between the two Research Data Centres of the Statistical Offices of the Federation and of the Länder, and together with subsequent users within the scientific community, who assist the work in the shape of a Advisory Board. FAST 98 is now available to scientists for Euro 65.00 via the Research Data Centres (www.forschungsdatenzentren.de).

1.3 The Advisory Board on the FAST 98 project

Anonymisation measures always entail reducing the information available in the existing individual data (information loss). Be it by means of blurring, deleting or falsifying, the resultant anonymised data always have limited potential for analysis in comparison to the original data material. What information is suppressed in the anonymisation process is at least partly variable. It is therefore possible to substitute information to a limited degree.

Since these substitution possibilities exist, there is a problem to be resolved as to deciding which information is to be suppressed in order to safeguard the confidentiality of the respondents in the data. For this reason, when creating FAST 98, this decision was discussed and decided on together with the subsequent users within a Advisory Board. The data of the income tax statistics have been available within the Statistical Offices for quite some time for scientific research, in particular in the framework of advice for the political arena. Information had therefore already been collected at the start of the project regarding a large group of subsequent users. These were written to and asked to help work on FAST 98. Both the Research Data Centres of the Statistical Office of the Federation and of the Länder considered that the project should be headed by someone outside official statistics, so that this position was put out to tender. The Advisory Board was finally composed of the following individuals and institutions:

Prof. Dr. Joachim Merz	University of Lüneburg (project leader)
Prof. Dr. Dr. Giacomo Corneo	Free University of Berlin
Dr. Markus Eltges	Federal Office for Building and Regional Planning
Prof. Dr. Heinz Galler	Martin Luther University of Tübingen
Mr. Hans-Joachim Georg	Bavarian Land Office for Statistics and Data Processing
Mr. Joachim Goletz	Land Office for Data Processing and Statistics of North Rhine-Westphalia
Mr. Volker Kordsmeyer	'Taxes' Division of the Federal Statistical Office

[9] cf. on this Zühlke, S.; Zwick, M.; Scharnhorst, S.; Wende, T.: The research data centres of the Federal Statistical Office and the statistical offices of the Länder in Schmollers Jahrbuch 4/2004 pp. 567-578, as well as www.forschungsdatenzentrum.de

Dr. Hermann Quinke	Fraunhofer Institute for Applied Information Technology
Dr. Claus Schäfer	Hans-Böckler Foundation
Prof. Dr. Viktor Steiner	German Institute for Economic Research
Dr. Stefan Weil	Land Statistical Office of Rhineland-Palatinate
Dr. Heike Wirth	Centre for Survey Research and Methodology
Dr. Sylvia Zühlke	Research data centre of the Land Statistical Offices
Mr. Markus Zwick	Research data centre of the Federal Statistical Office

At the first meeting, the fundamental procedure was discussed on the basis of an initial concept for anonymisation put forward by the Statistical Offices. On the basis of the results of this meeting, the first concrete de facto anonymised microdata file of the income tax statistics 1998 was created and extensively tested for sufficient anonymity of the respondents. The concept for anonymisation submitted for the second meeting was in turn intensively discussed and refined. This recently led to extensive work being carried out on drafting and testing. At the third meeting, which took place in April 2004, the concepts were accepted by the Advisory Board for the anonymisation of the income tax statistics 1998 and on the basis of the results of the confidentiality test that had been presented, the scientific-use file was categorised as de facto anonymous by the Statistical Offices and the participating lawyers.

Subsequently, the Advisory Board summed up as follows and made recommendations:[10]

"By creating a scientific-use file of income tax 1998, the informational infrastructure in Germany is sustainably improved. The income tax statistics are of considerable interest for the scientific community with regard to the differentiation of the income information, its quality as an official full survey, as well as its possibility to also describe maximum income.

FAST is a dynamic product. The practical experience of the scientific users working with it is being collected and incorporated in the scientific-use file of the income tax statistics 2001, which is the next to be developed, so that a methodical refinement is guaranteed. This also entails a permanent review of the degree of anonymisation required.

On the basis of the experience that has been collected, the Advisory Board is also in favour of developing a FAST regional file. Hence, FAST would be able in future to be supplemented with regional planning variables."

This article describes the approach followed and states reasons for the decisions leading to the de facto anonymous file. Whilst Chapter 2 explains in detail the income tax statistics and the 10% sample ob-

[10] cf. Statement of the Advisory Board on the FAST 98 project in the Annex

tained from it, Chapter 3 describes the conception for anonymisation of the income tax statistics 1998. In Chapter 4, you will find a description of the comprehensive tests for data protection. An outlook completes this essay.

2 Income tax statistics

2.1 Methodical basis and structure of the individual data of income tax statistics 1998[11]

Article 2 subsection 2 of the Act on Fiscal Statistics (Gesetz über Steuerstatistik - Steuerstatistikgesetz, StStatG) stipulates that the income tax statistics are to be collected every three years. Over and above this, it specifies the collection variables which are to be collected. These include, as well as the variables of the taxation process, socioeconomic variables, such as tax-payers' age or gender.

The income tax statistics are decentral, secondary statistics. This means that the information is not collected for the purpose of statistics, but is created in another context, in this case in the taxation process, and is used statistically at a second stage. To this end, the tax offices provide the respective information on the tax-payer to the Land Statistical Offices at set dates. The latter generate the respective results for the Länder and transmit the tables emerging from them to the Federal Statistical Office. The Federal Statistical Office then combines the Land results in the next step to create the Federal result. By reforming the Act on Fiscal Statistics in the context of the 1996 Annual Tax Act (Jahressteuergesetz)[12], in addition to the data contained in tables used to create a Federal result, the individual items of information provided by the Land Statistical Offices are also transmitted to the Federal Statistical Office, including for additional processing. This central availability of the individual data provides extensive analysis possibilities in the context of these statistics.

As secondary statistics, the income tax statistics depend on the income tax return implemented by the tax offices. Because of the periods granted to tax-payers to submit their income tax declaration, 2 3/4 years pass until the last data are available to the respective Land Statistical Offices. This therefore already causes a considerable time lag in the creation of statistical results. The possibility to accelerate the publication of the statistics by expanding initial results is made more difficult if large and complicated cases cannot be processed by the tax offices until the end of this period. For tax-payers with a high income in particular, for instance, there is an inherent incitement to extend the tax assessment to achieve interest advantages. The consequence of the three-year nature of the statistics and the periods for income tax return is therefore that it is only in the fourth year after the end of the assessment year in question that results are available, and these in some cases remain the most up-to-date until the seventh year. For instance, in 2004 the data on the assessment year 1998 are currently the most up-to-date of the income tax statistics.

Because of the varied nature of their data, the income tax statistics offer a large number of possibilities for analysis. Here, in addition to purely fiscal considerations, surveys may be implemented on the income spread. In particular those on high and highest incomes are not collected with this level of precision in any other statistical source than in the income tax statistics. This makes these statistics particularly valuable

[11] cf. concerning the information provided in this section Zwick, M.: Individual tax statistics data and their evaluation possibilities for the scientific community, in: Schmollers Jahrbuch, 2001, pp. 639 ff., as well as Rosinus, W.: Die steuerliche Einkommensverteilung, in Wirtschaft und Statistik, 6/2000, pp. 456-463.

[12] Reform of the "Act on Fiscal Statistics (Gesetz über Steuerstatistik" (StStatG)" with Article 35 of the 1996 Annual Tax Act of 11 October 1995 (Federal Law Gazette [BGBl.] Part I p. 1250) most recently amended by Article 56 of the Act of 23 December 2003 (Federal Law Gazette Part I p. 2848).

for an observation of this social group.[13]

When carrying out analyses, it must however be taken into account that the definitions of wage and income tax are based on fiscal law. For this reason, the variables cannot be simply compared with those from the national accounts. It is the term "total amount income" which is closest to the definition of income contained in the national accounts. However, this for instance only partly accommodates re-distributions, and is orientated more in line with tax-payers' primary market income. A household's actual available income is however influenced by State re-distributions, such as by the progressive income tax tariff or the transfer income which is only partly described in income tax statistics. In particular in distribution analyses, these restrictions must be taken into account. In the use of these data for distribution analyses, this has led to the convention that in the first step economic income should be calculated from the information of the income tax statistics [14].

The almost 30 million individual datasets of the income tax statistics 1998 encompass almost 500 variables per tax-payer, a different number of which are filled, depending on the tax case. The variables document the taxation process, starting with income through to the actual tax owed for each tax-payer.

It can be observed how net income before tax[15] is calculated (for instance gross income minus the income-related expenses). This is currently not possible with profit income[16], since the data contain no information on business receipts or expenses. Apart from via the limited-quality information contained in the Annex for statistical purposes, it hence remains impossible to trace how these types of income are created[17].

A data record represents a tax-payer. When married couples are assessed jointly, and the splitting system is applied, a tax-payer consists of two individuals or two tax cases. For this reason, the almost 30 million individual datasets comprise information on more than 42 million tax cases. Until the variable "gross income", the respective variables for the spouses are shown separately here. In the further course of taxation, this is no longer possible or no longer makes sense. As a consequence of the distinction made between tax-payers and tax cases, the fiscal income distribution based on the distribution of the "total amount income" is not a distribution of the individual income. However, it also does not precisely portray the distribution of household income since several tax-payers may live within a household (for instance assessed children living with their parents). Nevertheless, as a rule in the analyses the tax-payer is used as an approximation of the household and the income distribution is calculated on the basis of households.[18]

As was described at the start of this section, in addition to the quantitative variables of the taxation process, the datasets also show socioeconomic variables which facilitate a targeted analysis of individual groups within society. These variables include inter alia gender, regional distribution, religion, age and

---

[13] For instance in the context of the Wealth and Poverty Report of the Federal Government; cf. on this Merz, J. (2001), Hohe Einkommen, ihre Struktur und Verteilung – Mikroanalysen auf der Basis der Einkommensteuerstatistik; Lebenslagen in Deutschland - Der erste Armuts- und Reichtumsbericht der Bundesregierung, Federal Ministry of Health and Social Security, Berlin.

[14] cf. Bach, S., Bartholmai B.: Möglichkeiten zur Modellierung hoher Einkommen auf Grundlage der Einkommenssteuerstatistik, DIW – Diskussionspapiere No. 212, Berlin, 2000.

[15] Surplus income, calculated from cash receipts minus income-related expenses, includes income from dependent personal service, from rental and royalties, from investment of capital, as well as from other sources of income.

[16] Profit income, which is calculated from business receipts minus business expenses, includes income from agriculture and forestry, from trade, as well as from independent personal service.

[17] The Annex for statistical purposes is part of the obligatory information to be provided every three years on profit income recipients, but since the information is not needed for the taxation process, the tax offices have little motivation to adequately remind tax-payers to return the Annex for statistical purposes, not to mention to monitoring quality and plausibility.

[18] e.g. cf. Bach, S./Haan, P./Rudolph, H.-J./Steiner, V.: Reformkonzepte zur Einkommens- und Ertragsbesteuerung: Erhebliche Aufkommens- und Verteilungswirkungen, aber relativ geringe Effekte auf das Arbeitsangebot, in: DIW-Wochenbericht 16/2004, as well as Rosinus, W.: Die steuerliche Einkommensverteilung, in Wirtschaft und Statistik, 6/2000, pp. 456-463

with entrepreneurs the economic activity ('trade code', GKZ93).

2.2 The representative sample of income tax statistics

The statutory basis for the sample from the income tax statistics is to be found in article 7 subsection 4 of the Act on Fiscal Statistics, and is prescribed inter alia as a 10% sample. It serves to estimate the financial and organisational impact of amendments of regulations as the fiscal and transfer system is refined.[19] The sample plans for this are worked out centrally in the Federal Statistical Office. The samples are taken from the individual data of the overall material transmitted by the Land Statistical Offices. The sample was created in 1998 and in previous years as a stratified random sample. A high precision requirement applied here as a selection criterion, in particular as to the documentation of the total amount income.

The Act on Fiscal Statistics requires a sample which is "nationally representative". For this reason, the Federal Land was abandoned in the samples for the assessment years 1992 and 1995 as a variable for stratifying, and stratifying only by old and new Federal Länder was implemented.[20] In order to make Land analyses possible that are as exact as possible, in the sample for the assessment year 1998 the Federal Land was included as a variable for stratifying. Overview 1 compares the respective variables for layering of the samples taken for assessment years 1992, 1995 and 1998.

Overview 1: Variables for stratifying of the income tax statistics

1992		1995		1998	
Variable	Categories	Variable	Categories	Variable	Categories
Federal Land new/old	2	Federal Land new/old	2	Federal Land	16
Type of assessment	4	Type of assessment	4	Type of assessment	2
Child allowance steps	4	Child allowances	4	Children	3
Primary type of income	7	Primary type of income	7	Primary type of income	3
Total amount income	7	Total amount income	12	Total amount income	7

Whilst the variables for stratifying basically remained the same, the respective numbers of the categories changed. This applies in particular to the most up-to-date sample of 1998. To the 2,016 strata which emerged from the complete combinations of the variables, another 32 strata are added for 1998. These consist of the so-called manual cases (wage-tax cards which are not assessed). With these, stratifying was implemented only by the 16 Federal Länder, as well as by two income classes. In total, the number of strata in 1998, at 2,048, is lower than the 1995 number (2,704 layers, incl. "special strata"). In 1992, only 1,344 strata were formed.[21]

Strata with few entries are as a rule contained in the sample as a full survey. Furthermore, all respondents with a total amount income higher than Euro 102,257 (DM 200,000) were fully included in the sample because of their heterogeneity.

[19] Article 7 subsection 4 of the Act on Fiscal Statistics.
[20] cf. Zwick, M.: Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken, in Wirtschaft und Statistik, 7/1998, p. 570.
[21] cf. Zwick, M.: Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken, in: Wirtschaft und Statistik, Vol. 7, 1998, pp. 570.

3 Anonymisation concept for the income tax statistics 1998

The 10%-sample that is described at 2.2 was used as a data basis for the anonymisation of the income tax statistics 1998. The principle of disproportionality described in the introduction is only a necessary precondition for a scientific-use file. This precondition guarantees the de facto anonymity of data, but not that the data can be used for scientific analysis. It would make sense for de facto anonymous data to be provided to the scientific community by the Statistical Offices as a scientific-use file only if they offer sufficient scientific analysis possibilities. Since anonymisation of respondents always implies a reduction of information, it follows that anonymisation is to be restricted to the necessary degree. In order to achieve this, in FAST 98 the respondents were anonymised to a degree corresponding to their re-identification risk. Those with a lower re-identification risk are subject to a lower degree of anonymisation than those with higher risks. What is more, data-altering anonymisation procedures, as tested above all with individual items of economic statistical data, were not used[22]. Only procedures were used which have already been used for some time in other personal individual data within the Statistical Offices.[23,24]

3.1 The Christmas tree anonymisation principle

Not every one of the roughly 2.8 million respondents of the sample could be individually tested for its re-identification risk. Rather, it was presumed that the risk of re-identification increases in line with the amount of income. On the basis of this presumption, therefore, the respondents were divided into a variety of income ranges and labelled with an indicator for their risk. Within the anonymisation ranges, anonymisation methods were carried out that were specifically adjusted to the risk (cf. Overview 2). Analogously to the Christmas tree, which shows less green as the trunk becomes higher, the data show less information as income increases because of the anonymisation methods (Christmas tree anonymisation).

With the aid of the total amount income, the data with the positive income were sub-divided in five ranges (cf. Overview 2). The first covers a total amount income from zero to twice the average total amount income. The second range goes from this to the 99% percentile of the income distribution. The third range covers the interval from the 99% percentile to the 99.95% percentile, whilst the fourth range covers this threshold until the 1,000 respondents which show the highest total amount income. The fifth range is formed by the 1,000 individuals with the highest total amount income. In the cases in which there is no total amount income, the "total gross wages" were used as a sub-division variable.

[22] As to the greater impact of data-altering procedures in comparison to traditional anonymisation cf. Rosemann, M./Vorgrimler, D./Lenz, R.: Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Allgemeines Statistical Archiv, Vol. 1, 2004, pp. 73-99.

[23] A general overview of anonymisation methods can be found in Höhne, J.: Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Gnoss, R./Ronning, G. (eds.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik, 2003, pp. 69-94, as well as on anonymisation in Federation statistics cf. Köhler, S.: Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: Forum der Bundesstatistik Vol. 31, 1999, pp. 133-150.

[24] An exception is that of the three tax-payers with the highest income, whose variables were additionally anonymised using microaggregation, as described below.

Overview 2: Sub-division of the anonymisation ranges

Anonymisation range	Positive total amount income in EUR (DM)	Negative total amount income in EUR (DM)
1	0 to 64,106 (0 to 125,381) (double the average total amount income)	0 to – 102,258 (0 to – 200,000)
2	64,107 to 137,532 (125,382 to 268,990) (99% percentile)	-
3	137,533 to 970,202 (268,991 to 1,897,552) (99.95% percentile)	– 102,259 to – 511,292 (– 200,001 to – 1,000,000)
4	970,203 to 7,354,714 (1,897,553 to 14,384,571) (up to the 1,000 richest)	-
5	› 7,354,714 (› 14,384,572)	– 511,292 (– 1,000,000)

The "income" variable was used to characterise those tax-payers which have negative income, in the event of the total amount income not being filled. To anonymise the respondents with negative income, three ranges were formed (cf. Overview 2). The first range covers those respondents whose negative income lies between –1 and the 95% percentile of the absolute negative income distribution. The second range stretches from this boundary up to the 99.5% percentile, whilst the third range comprises all remaining respondents with the absolutely highest negative incomes. Absolute boundaries have still been used in the current version instead of these relative ones (cf. Overview 2). The anonymisation methods are identical to the methods applied in ranges one, three and five of the respondents with positive income.

Table 1 shows the significance of the anonymisation ranges according to the criteria tax-payer, total amount income and income tax determined. It shows that anonymisation range 1 is the most significant by far with regard to tax-payers. This however reduces somewhat with the value observations, which are connected to the skewness of the income distribution.

Table 1: Distribution of the anonymisation ranges (in percent)

Anonymisation range	Tax-payers		Total amount income		Income tax	
	share	cumulated	share	cumulated	share	cumulated
1	92.2	92.2	68.8	68.8	51.7	51.7
2	6.6	98.8	17.1	85.9	22.2	73.9
3	1.2	99.99	8.6	94.5	16.2	90.1
4	0.05	99.99	3.2	97.7	6.2	96.3
5	0.02	100	2.3	100	3.7	100

3.2 General anonymisation

In addition to anonymisation adjusted to the amount of income, methods were taken with which all respondents were at least anonymised (general anonymisation). Overview 3 provides information on these anonymisation methods.

Overview 3: General anonymisation methods

Entry field	Variable(s)	Methods
EF1	Reason for assessment	Re-coding of the eight attributes in: 1 = assessed cases 2 = manual cases
EF13 + EF14	Religions (in each case separated for men and women)	Re-coding of the twelve attributes in: 1 = Protestant 2 = Catholic 3 = others 4 = none
EF19	Type of assessment	Re-coding of the eight attributes in: 1 = basic table 2 = splitting table
EF64 + EF67	Age (in each case separated for men and women)	Introduction of a lower limit (15 years) and an upper limit (70 years). Above or below the limits, the age was stated as the average of those above or below the limits.
c36010 – c37066	Number. of children	The variables of the children were removed. Only the number and information on the age of the children are contained in the data. 5 and more children were assigned to the attribute ‹ children.

The restriction of the income tax data to a 10% sample, moreover, constitutes a general anonymisation methods since because of the sample a potential data intruder has no knowledge of whether the specific respondents is contained in the sample at all.[25] This makes successful re-identification more difficult and less certain. However, the sample was not taken to be anonymised, but with the goal of obtaining "manageable" volumes of data with the highest possible representativeness.[26] For this reason, smaller heterogeneous groups of respondents are contained as complete surveys. As has been seen, this applies to those earning higher incomes in particular. For these, a potential data intruder hence has knowledge of participation, so that the above argument does not apply. Therefore the sample has its greatest impact as to anonymisation in the range of low and medium incomes.

The age of the data has a two-fold impact as an anonymisation methods. Firstly, it is more difficult for a potential data intruder to generate relevant additional knowledge for a respondents the older the data are. Secondly, the usefulness of an item of information is contingent on its topicality. For this reason, the usefulness of identification reduces as it gets older. This argument however only applies if the timeliness of the data is a positive element of the untility function for the potential data intruder.

[25] On knowledge of participation cf. also Lenz, R./Sturm, R./Vorgrimler, D.: Maße für die faktische Anonymität von Mikrodaten, in: Wirtschaft und Statistik, Vol. 6, 2004, pp. 623-624.
[26] On the function of the sample cf. Zwick, M.: Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistik, in: Wirtschaft und Statistik, in: Wirtschaft und Statistik, Vol. 7, 1998, pp. 566-572.

3.3 Specific anonymisation

3.3.1 Categories of variables

In the anonymisation ranges described, variables were differently blurred or deleted. To achieve this, the continuous variableswere sub-divided into three categories as to their significance. The first contains the variables which are also shown in the respondents with the highest incomes. The second category contains variables which are only dealt with for the highest incomes, whilst the variables of the third category are restricted first.

Variables of the first category:

- gross income (A+B[27])
- total amount income
- earnings
- taxable income
- income tax according to the basic scale
- assessable income tax

Variables of the second category:

- income from agriculture and forestry (A+B)
- income from trade (A+B)
- income from independent personal service (A+B)
- income from dependent personal service (A+B)
- income from investment of capital (A+B)
- income from rentals and royalties (A+B)
- other income (A+B)
- special expenses which are not expenses of a provident nature
- special expenses: expenses of a provident nature
- extraordinary financial expense, deductible – with separate assessment –A –
- extraordinary financial expense, deductible – with separate assessment –B –
- incentives for home ownership: total tax concessions

All other roughly 300 continuous variables belong to the third category.

Information which for purposes of anonymisation is either only blurred, falsified or no longer contained in the target data at all is less valuable to a data intruder than the original information.[28] Anonymisation therefore impacts not only the cost incurred by a data intruder, but also the benefit is negatively influenced. In the income tax statistics, this aspect applies to the continuous variables in particular. These may be difficult to use as key variables, meaning that changing their values does not provide any additional protection to the respondents, but the continuous variables are likely to be of most use to a data intruder. If therefore continuous variables are deleted from the data or blurred, this has an effect above all on the benefit side of the data intrusion. This is a major aspect to achieve de facto anonymity in which the disproportionality of a data intrusion is also accommodated.

[27] A are male, B female tax-payers.

[28] cf. on this Höhne, Sturm, Vorgrimler: Konzept zur Schutzwirkung faktischer Anonymisierung, in Wirtschaft und Statistik, Vol. 4, 2003, pp. 287-292 as well as Lenz, R./Sturm, R./Vorgrimler, D.: Maße für die faktische Anonymität von Mikrodaten, in: Wirtschaft und Statistik, Vol. 6, 2004, pp. 621-638.

3.3.2 Special anonymisation methods

Overview 4 summarises the special anonymisation measures taken for the different sub-ranges.

Overview 4: Special anonymisation methods in the income ranges

Variable		Anonymisation range[1)]				
		1	2	3	4	5
Religion		4 attributes	4 attributes	not stated	not stated	not stated
Children		No. up to four Age of first three children	No. up to four Age as dummy	No. up to four Age as dummy	No. up to four	Yes/No
Age		Yes with 15 / 70 limit	Class with 5 years	Class with 10 years	Class with 10 years	Class with 10 years
Region		Federal Land	Federal Land	West/East	West/East	West/East
Trade code		1-digit	1-digit	1-digit	1-digit	not stated
Freelancers		9 attributes	9 attributes	9 attributes	9 attributes	Dummy yes /no
continuous variables	1	Yes	Yes	Yes	Yes	Yes
	2	Yes	Yes	Yes	Yes, but male female as total	Dummy (+ significance variables)
	3	Yes	Yes	Yes	Dummy	No

1) With positive income: 1 = from 0 to a total amount income of Euro 64,106; 2 = 64,107 to 137 532 EUR (99% percentile); 3 = 137,533 to Euro 970,202 (99.95% percentile); 4 = Euro 970,203 to the 1,000 highest sums of all income; 5 = the 1,000 highest sums of all income + Members of Parliament. With negative income: 1 = from 0 to a negative income of Euro 102,258 (95% percentile); 3 = from 102,259 to a negative income of Euro 511,292 EUR (99.5% percentile) EUR; 5 = with a negative income over Euro 511,292.

The scientists involved in the discussion of the anonymisation concept prefer to conserve the continuous variables as against the socioeconomically discrete variables. This is reflected in the special anonymisation in that firstly the discrete variables were blurred or deleted before the continuous variables were taken for anonymisation. For instance, the variable "Religion" is only represented with four possibilities in the first two ranges, and only in these ranges is "Region" connected with the Federal Land. Age is given a class from the second range and in the first, an upper and lower boundary was imposed.[29] The number of children is no longer stated in the fifth range. By contrast, in the first range not only the number, but also the age of the first three children is included in the data.

Because of this more incisive encroachment on the socioeconomic variables, the continuous variables up to and including the third range could remain unchanged in the data. The third category still contains the variables of the fourth range as a dummy variable, whilst they are deleted in the fifth. Transforming the information into a dummy variable means in this context that the new (dummy) variable takes on a '1' as an attribute if the original variable was available with a positive attribute, a zero shows when the original variable was not available from the tax-payer, and a –1 indicates negative original values. The variables of the second category are still shown in the fourth range, but with no gender-specific division. These variables are still contained in the fifth as a dummy variable. There is only one restriction with the variables of the first category. The values of the three respondents with the highest attributes in each case were re-

[29] The age of those above and below the boundary is given as the average age of those above or below the boundary.

placed by the average values of their respective attributes (microaggregation[30]). Thus, the maximum values of the variables of the first category no longer correspond to the original values, but show the arithmetic mean of the three highest values, whilst what is conserved is the full total of the value.

3.3.3 Additional Informations of the anonymised file

In addition to reducing information through anonymisation, additional generated information has been included in the file which is to provide supplemental valuable information to the scientific community.

For tax-payers with income from free professions, the variable "freelancer" was generated from the trade code available from the original income tax statistics with the following attributes in the first four anonymisation ranges:

Architectural and engineering activities and related technical consultancy, research; law offices with notaries public; auditing activities, economic advisers; general practitioners; other health professions; advertising, activities of the photographic industry, art and culture; literary creation; schools and others.

In addition, the data contain a dummy variable which states whether the tax-payer works on a freelance basis. Hence, one follows a long tradition of fiscal statistics in which the group of the independent professions is evaluated and analysed in this standardisation process.

Anonymisation range 5 only contains the variables of the second category as a dummy variable. So that the data users can also imitate the structure of incomes in the highest income range, the seven types of income were sub-divided into three categories (profit income, income from dependent personal service and other net income before tax). Each of these categories has a significance variable. This takes on the value 1 if the highest income is made in this type of income, and 3 if the lowest income comes from this category, correspondingly this variable shows the attribute 2 for a medium significance. If no income stems from the category, the variable is set to 0. As an example, Table 2 states the frequency distribution of the variables for the 1,000 respondents with the highest incomes. Accordingly, it shows which income categories contribute most towards reaching the highest income.

Table 2: Income structure for the 1.000 taxpayers with the highest income

Significance	Profit income	Income from employment	Other net income before tax
High	910	10	80
Medium	49	318	616
Low	30	275	283
None	11	397	21

The anonymisation amount of each respondent is stated as further additional information. attributes 1-5 reflect the anonymisation ranges here. Additionally, an attribute 6 was introduced which refers to those respondents whose continuous variables were gradually microaggregated.

[30] On microaggregation cf. Domnigo-Ferrer, J./Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control, IEE Transaction on Knowledge and Data Engineering, Vol. 14(1), 2002, pp. 189 et seqq.

4 Test of data protection

The test of data protection uses the anonymised data described in Chapter 3. By nature, the possibilities of re-identification are greater in the original data. This portrayal is not used here since these data because of the ex ante appraisals of the deanonymisation risk were not considered for a scientific-use file.

4.1 Approaches that might be taken in re-identification attempts

Before an anonymised file can be considered de facto anonymous within the meaning of article 16 subsection 6 of the Federal Statistics Law, it must be examined for sufficient data protection.[31] A possibility to do so is offered by simulations of re-identification attempts. These can be sub-divided into two types: So-called mass attack aims to use external databases as additional knowledge to re-identify as many respondents as possible, whilst in individual intrusions an attempt is made to find a specific respondent in the anonymised data. Both procedures have been used in the past to examine anonymised data as to their protection.[32]

As a first step in testing the adequacy of anonymisation, on principle one must examine which of the two procedures can be considered for a protection test on the income tax statistics 1998.

In order to be able to successfully re-identify respondents of an anonymised file, the following fundamental presumptions for a data intruder are required:[33]

- additional outside knowledge of the specific respondent (such as in the shape of a database),
- knowledge of participation in the survey by the specific respondent,
- variables which are contained both in external and in target data (key variables)

These conditions considerably restrict the possibilities to which re-identification attempts are amenable. Mass attack appears to be ruled out in fact since the additional knowledge required shows neither the necessary key variables, nor does it exist in a suitable and adequate form.[34] For individual intrusions with specific groups of individuals, however, sufficient additional information is available, even if the additional knowledge must be gathered from a variety of sources. The main attention here must be paid to those individuals who are contained in the sample as a complete survey because of their special status. Only for this group does a data intruder have knowledge of participants. Since those on medium and low incomes as a rule are only contained as a sample, these can be regarded ex ante as being de facto anonymous, even if more information is available for these because of the lower level of anonymisation.[35] Respondents who are individually entered in the sample can be recognised by virtue of the fact that their expansion factor takes on the value 1. Because of these arguments, the protection analysis focuses on individual intrusions on the following groups:

---

[31] More information on testing for data security of the anonymised income tax statistics 1998 cf. Scharnhorst, S./Zühlke, S./Stegenwaller, L.: Beiträge zum Projekt "Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998", appears in the series DZ-Arbeitspapiere, www.forschungsdatenzentrum.de.

[32] For instance for the cost structure survey in manufacturing cf. Lenz, R.: Disclosure of confidential information by means of multi-objective optimisation. Proceedings of the Comparative Analysis of (micro) Enterprise Data Conference (CAED), London 2003. Internet: http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp and Vorgrimler, D.: Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios, in: Gnoss, R./Ronning, G. (eds.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik, 2003, pp. 40-59. For turnover tax statistics Lenz, R./Vorgrimler, D.: MAtching der German Turnover Tax Statistics, in FDZ-Arbeitspapiere Vol. 4, 2005.

[33] Brand, R., Bender, S., Kohaut, S.: Possibilities for the creation of a scientific-use-file for the IAB-Establishment-Panel, Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Protection, Thessaloniki (1999) 57-74

[34] cf. Scharnhorst, S./Zühlke, S./Stegenwaller, L.: Beiträge zum Projekt "Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998", appears in the series FDZ-Arbeitspapiere, www.forschungsdatenzentrum.de.

[35] cf. Scharnhorst, S./Zühlke, S./Stegenwaller, L.: Beiträge zum Projekt "Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998", appears in the series FDZ-Arbeitspapiere, www.forschungsdatenzentrum.de.

- famous personalities, company executives,
- individuals with freelance activity,
- Members of Parliament, and
- personal environment.

4.2 Key variables

Borrowing from Elliot/Dale, it is possible for the key variables from the additional knowledge to be subdivided into the following four categories:[36]

1. high-quality easily-accessible additional knowledge (prime keys)
2. low-quality easily-accessible additional knowledge (background keys)
3. hard-to-access high-quality additional knowledge (critical keys)
4. hard-to-access low-quality additional knowledge (inefficient keys)

Here, the quality of the additional knowledge depends on how strongly it distinguishes the data, how stable the additional knowledge is over time, and how high is the probability of measurement errors.

In the income tax statistics, the age of the tax-payers and the number of children can be regarded as *prime keys*. For specific groups of individual, such as freelancers, the information on their belonging to a professional group is also a prime key. The information on their place of residence and gender could be used as *background keys* (easy to find out, but only with a low differentiating impact). The *critical keys* include information on the age of the first three children, religion, donation activity and maintenance obligations.[37]

4.3 Results of re-identification attempts

- **Famous people and company executives**

Famous people were observed both from the world of sport, and from media. The risk of re-identification was considered to be high from the outset, especially among sportspersons, since they more frequently show the relatively rare combination of variables high income / low age. However, in none of the total of 12 individual intrusions on famous people could a clear attribution be achieved. Furthermore, the addition of other key variables – if they had been available – does not cause re-identification to become more probable.

One comes to the same result in the range of economic managers, although more income information is available for this group of individuals in the additional knowledge. Here too, none of the six target persons could be re-identified, despite intensive research, in particular via the Internet.

On the basis of the results with a total of 18 famous personalities and company executives, the individual intrusions on these groups of individuals can be regarded as having failed.

- **Individuals with freelance activity**

[36] cf. Elliot, M./Dale; A.: Scenarios of attack: the data intruder's perspective on statistical disclosure risk. In: Netherlands Official Statistics, pp. 6-10

[37] For detailed reasoning on this sub-division cf. Scharnhorst, S./Zühlke, S./Stegenwaller, L.: Beiträge zum Projekt "Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998", appears in the series FDZ-Arbeitspapiere, www.forschungsdatenzentrum.de.

As explained in section 4.2, the variable of freelance activity of a respondent can be used as a prime key since this characterises tax-payers as a part of a relatively small sub-group.[38] The variable of freelance activity has been included as additional information in the de facto anonymised data, thus making it easier to use as an overlap variable (cf. section 3.2.3). For this reason, "freelancers" were subjected to a separate test. Having said that, this did not consist of "real" individual intrusions, instead the share of one-off attribute combinations in the data was ascertained. For North Rhine-Westphalia, this was so low that no endangerment to data protection is to be feared. For Mecklenburg-Western Pomerania, the proportion is much higher, but such precise information on the respondent is needed in order in reality to identify a one-off case that this appears to be possible only in a freelancer's personal environment. The sole fact of being in the group of freelancers, by contrast, does not endanger the security of the respondents.

- **Members of Parliament**

The re-identification of Members of Parliament is made easier both by improved access to the additional knowledge, and by better information in the target data. The amount of the allowances received by Members of Parliament is publicly accessible, and is contained as a separate variable in the data on income tax statistics. In the first step of anonymisation, it was already manifest that this variable must be combined with at least one more variable within the same type of income (other income). Real individual intrusions were carried out on the basis of these measures.

In an initial wave of tests, a search was carried out in the data for 16 Members of Parliament. Clear and correct re-identification was however only successful in one case.

The second step consisted of a change in the search direction. Since individual attributes of the variable "other income from performance" were more common with values which correspond to the daily expense allowance of the North Rhine-Westphalia Land Parliament and Federal Parliament Members, it was possible to identify a group of 86 Members from North Rhine-Westphalia. Eleven of these individuals could be clearly attributed. Over and above this, another two duplicate attributions are present. Of this total of 15 attributions, ten can be considered correct.

Since it appears to be possible with a relatively slight effort to identify Members of Parliament despite the anonymisation so far carried out, anonymisation should be tightened up for this sub-population. To this end, it was necessary to eliminate the identifying information. Not only was it possible for Members to be attributed to this group via frequency evaluations of the variable "other income from performance", but over and above this they could be placed in their individual Federal Länder. Further tests showed that this is possible as a rule as long as the income type "other income" is shown. It is only in anonymisation range 5 that this variable is only contained as a dummy variable. For this reason, as additional protective measures all respondents identified as Members of Parliament have been assigned anonymisation range 5. For this reason, anonymisation range 5 contains not only the 1,000 respondents with the highest positive total amount income and roughly 2,200 with the highest negative total amount income, but in addition roughly 3,000 Members of Parliament. The tightened up anonymisation measures of anonymisation range 5 offer additional protection, so that after this review of anonymisation the sub-population of Members of Parliament can also be regarded as being de facto anonymous.

[38] In the FAST data, roughly 260,000 datasets are labelled as male and roughly 100,000 datasets as female tax-payers with freelance income. Expanded for Germany, these correspond to roughly 800,000 male and roughly 400,000 female tax cases with freelance income.

- **Personal environment**

It is in general terms for the sub-group "personal environment" that the best additional knowledge available can be presumed for a data intruder. This applies both from a quantitative (number of key variables available) and from a qualitative point of view (reliability of the values).

Three individuals from the personal environment were looked for as target persons in the de facto anonymised data. With the first two individuals, a successful search was already unlikely since the target persons were attributed to strata with very low sampling fraction. The attempt at re-identification was hence not expedient for this reason alone.

With the third person, a clear attribution could be achieved which however turned out to be incorrect on examination. For this reason, the re-identification attempts in the personal environment could also be regarded as having failed. It should be pointed out that the personal environment of a data intruder in particular is a highly-subjective assessment. It is hence even more difficult to arrive at an "objective" evaluation than with the other sub-populations of the dataset.

4.4 Conclusion of the re-identification attempts

The re-identification attempts carried out on the basis of individual intrusions have shown that the respondents are de facto anonymous. This conclusion was reached unanimously by the Statistical Offices involved, the lawyers consulted and the advising scientific group of users. De facto anonymity could however only be ensured for Members of Parliament through additional anonymisation methods.

This result does not rule out a data intruder theoretically re-identifying a respondent. Having said that, absolute anonymity is not demanded by the legislator in article 16 subsection 6 of the FSL, but the lengths that a data intruder would have to go to for successful re-identification must be disproportionately high. This precondition is met, and hence the data are de facto anonymous and can be transmitted to the scientific community subject to the further strict conditions of article 16 subsection 6 of the FSL.

5 Outlook

With the de facto anonymised microdata file of the income tax statistics 1998, the official statistics are expanding the standardised scientific-use files offered. The Research Data Centres of the Statistical Offices of the Federation and of the Länder are hence meeting their obligation in particular to make research possible in situ, in addition to providing more ways of accessing data.

The scientific-use file is available to the scientific community for Euro 65 via the Research Data Centres of the Statistical Offices of the Federation and of the Länder. The low price is a result of the support of the Federal Ministry of Education and Research. The financial subsidy obtained from the political arena enables official statistics to provide the Research Data Centres with capacities permitting them to implement anonymisation projects and to offer the resultant scientific-use files at an affordable price.

However, the data requirement of the scientific community is not covered by standardised de facto anonymous microdata which can be created by the range of official statistics within the limits imposed by article 16 subsection 6 of the FSL. What is left is research reports which cannot be sufficiently researched with FAST 98. Thus, detailed surveys on high income or analyses with a detailed regional sub-division will only be possible to a restricted degree because of the reduction in information by the anonymisation measures.

In this case, the additional access routes via the Research Data Centres offer paths towards a solution. In addition to standardised (off-site) scientific-use files, (on-site) scientific-use files can be used in the individual ranges of official statistics that are created for the purpose of research at Safe Scientific Workstations. Furthermore, the possibility exists to use the complete information potential of official individual data via the controlled remote computers.[39] The declared goal of the Research Data Centres of the Statistical Offices of the Federation and of the Länder is to no longer permit research projects to fail because of a lack of access to official individual data. It will not always be possible to offer the most convenient path to official data, but it should always be possible to use the information potential of individual official data for the scientific community at a reasonable price. Costs are incurred however, for instance as information is suppressed or greater distances need to be covered to obtain information as a result of the data protection requirements. Official statistics must also respect these requirements, as equally they must adhere to freedom of research.

The justified interest of the scientific community in data that are as up-to-date as possible requires that income tax statistics cannot be completely anonymised in FAST 98. Rather, it is now a matter of building on the work done to offer updated scientific-use files. As a first step, the data of the assessment year 2001 are to be de facto anonymised as soon as they are available. Beyond this, it is a matter of examining whether income tax data anonymised after 2001 could in fact be offered on an annual basis. The data collected annually on income tax statistics could serve as a basis in accordance with article 2a of the Act on Fiscal Statistics. This would optimise the degree of topicality and comparability between the assessment years.

With the scientific-use file on the income tax statistics 1998, herewith presented, and the projects that have been launched for the anonymisation of salary and wage structure statistics, hospital statistics and the project result to be submitted in the summer of 2005 for the anonymisation of economic statistical data, the Statistical Offices have taken a key step towards improving the informational infrastructure. Hence, official statistics are able to balance out to some degree the constitutional conflict of interests between 'freedom of research' versus 'data protection'.

[39] cf. on this Zühlke, S.; Zwick, M.; Scharnhorst, S.; Wende, T.: The research data centres of the Federal Statistical Office and the statistical offices of the Länder in Schmollers Jahrbuch 4/2004 pp. 567-578 as well as www.forschungsdatenzentrum.de

Literature

BACH, S., BARTHOLMAI, B.: Möglichkeiten zur Modellierung hoher Einkommen auf Grundlage der Einkommenssteuerstatistik DIW – Diskussionspapiere No. 212, Berlin 2000.

BACH, S., HAAN, P., RUDOLPH, H.-J., STEINER, V.: Reformkonzepte zur Einkommens- und Ertragsbesteuerung: Erhebliche Aufkommens- und Verteilungswirkungen, aber relativ geringe Effekte auf das Arbeitsangebot, in: DIW-Wochenbericht 16, 2004.

BRAND, R., BENDER, S., KOHAUT, S.: Possibilities for the creation of a scientific-use-file for the IAB-Establishment-Panel, Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Protection, Thessaloniki (1999) 57-74

COMMISSION FOR THE IMPROVEMENT OF THE INFORMATIONAL INFRASTRUCTURE BETWEEN THE SCIENTIFIC COMMUNITY AND STATISTICS (ED.): "Wege zu einer besseren informationellen Infrastruktur", Baden-Baden, 2001.

ELLIOT, M., DALE; A.: Scenarios of attack: the data intruder's perspective on statistical disclosure risk, in: Netherlands Official Statistics, pp. 6-10, 1999.

DOMNIGO-FERRER, J., MATEO-SANZ, J.M.: Practical data-oriented microaggregation for statistical disclosure control, IEE Transaction on Knowledge and Data Engineering, Vol. 14(1), 2002, pp. 189.

HAUSER, R., WAGNER, G., ZIMMERMANN, K.: Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung: Ein Memorandum; Allgemeines Statistical Archiv, Vol. 82, 1998, pp. 369-379.

HÖHNE, J., STURM, R., VORGRIMLER, D.: Konzept zur Schutzwirkung faktischer Anonymisierung, in: Wirtschaft und Statistik, Vol. 4, 2003, pp. 287.

KÖHLER, S.: Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: Forum der Bundesstatistik Vol. 31, 1999, pp. 133-150

LENZ, R.: Disclosure of confidential information by means of multi-objective optimisation. Proceedings of the Comparative Analysis of (micro) Enterprise Data Conference (CAED), London 2003. Internet: http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp

LENZ, R., VORGRIMLER, D.: Matching the German Turnover Tax Statistics, in: FDZ-Arbeitspapiere Vol. 4, 2005, Internet: http://www.forschungsdatennetzwerk.de/veroeffentlichungen.asp

LENZ, R., STURM, R., VORGRIMLER, D.: Maße für die faktische Anonymität von Mikrodaten, in: Wirtschaft und Statistik, Vol. 6, 2004, pp. 623-624.

MERZ, J. Hohe Einkommen, ihre Struktur und Verteilung – Mikroanalysen auf der Basis der Einkommensteuerstatistik; Lebenslagen in Deutschland - Der erste Armuts- und Reichtumsbericht der Bundesregierung, Federal Ministry of Health and Social Security, Berlin, 2001.

MÜLLER, W., BLIEN, U., KNOCHE, P. AND WIRTH, H.: Die faktische Anonymität von Mikrodaten; Vol. 19 of the series of publications entitled Forum der Bundesstatistik, Federal Statistical Office, 1991.

ROSEMANN, M., VORGRIMLER, D., LENZ, R.: Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Allgemeines Statistical Archiv, Vol. 1, 2004, pp. 73-99

ROSINUS, W.: Die steuerliche Einkommensverteilung, in: Wirtschaft und Statistik, Vol. 6, 2000, pp. 456-463.

SCHARNHORST, S., ZÜHLKE, S., STEGENWALLER, L.: Beiträge zum Projekt "Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998", appears in the series FDZ-Arbeitspapiere, http://www.forschungsdatenzentrum.de

VORGRIMLER, D.: Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios, in: Gnoss, R./Ronning, G. (eds.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik, 2003, pp. 40-59.

ZÜHLKE, S., ZWICK, M., SCHARNHORST, S., WENDE, T.: The research data centres of the Federal Statistical Office and the statistical offices of the Länder in: Schmollers Jahrbuch Vol. 4, 2004 pp. 567-578

ZWICK, M.: Individual tax statistics data and their evaluation possibilities for the scientific community, in: Schmollers Jahrbuch, Vol. 4, 2001, pp. 639 ff

ZWICK, M.: Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken, in: Wirtschaft und Statistik, Vol. 7, 1998, pp. 566.

Statement of the Advisory Board

The submission of the final report of the Commission for the Improvement of the Informational Infrastructure between the Scientific Community and Statistics has once more underlined the desire of the scientific community for further de facto anonymised data stocks from the field of official statistics. Thus, the Commission recommends amongst other things " . . . to promote the development of scientific-use files (SUF) as a major tool for micro access".

The Research Data Centres of the Statistical Offices of the Federation and of the Länder, which were the product of another recommendation by the Commission, have tackled this topic, and together with the specialist statisticians and the scientific community have created a de facto anonymised microdata file on income tax statistics (FAST). The work on FAST was assisted by the above groups within a Advisory Board.

The Advisory Board was composed of

– the project leader (Prof. Dr. Merz),

– the "Taxes" Division of the Federal Statistical Office,

– the Research Data Centres of the Statistical Offices of the Federation and of the Länder,

– the "Taxes" Specialist Divisions of the Land Statistical Offices of North Rhine-Westphalia, Rhineland-Palatinate and Bavaria,

- as well as by representatives of the group of scientific users.

The job to be done

The task of the Advisory Board as a joint body of data producers and users was to assist with advice and evaluation in the creation of FAST. Within this body, firstly various anonymisation concepts were the subject of intensive discussion from the point of view of data protection, as were also the analysability of the anonymised file once it had been created.

The result

On the basis of the concepts that have been developed for anonymisation of the income tax statistics 1998, as well as of the intrusion scenarios that have been tested on the anonymised file, after intensive discussion within the Advisory Board a scientific-use file of the income tax statistics 1998 was developed which in the view of the Advisory Board is de facto anonymous within the meaning of article 16 subsection 6 of the Federal Statistics Law.

Scientific research hence has data material at its disposal in the shape of the scientific-use file of the income tax statistics 1998 making it possible to answer a large number of fiscal policy questions. What is more, FAST makes it possible to tackle socioeconomic questions such as those in the context of the second Wealth and Poverty Report of the Federal Government 2004.