Merz, Joachim; Vorgrimler, Daniel; Zwick, Markus

Article

De Facto Anonymised Microdata File on Income Tax Statistics 1998

Schmollers Jahrbuch – Zeitschrift für Wirtschafts- und Sozialwissenschaften. Journal of Applied Social Science Studies

Provided in Cooperation with:
Duncker & Humblot, Berlin

Suggested Citation: Merz, Joachim; Vorgrimler, Daniel; Zwick, Markus (2006) : De Facto Anonymised Microdata File on Income Tax Statistics 1998, Schmollers Jahrbuch – Zeitschrift für Wirtschafts- und Sozialwissenschaften. Journal of Applied Social Science Studies, ISSN 1865-5742, Duncker & Humblot, Berlin, Vol. 126, Iss. 2, pp. 313-327, https://doi.org/10.3790/schm.126.2.313

This Version is available at:
https://hdl.handle.net/10419/292159

Schmollers Jahrbuch 126 (2006), 313 – 327
Duncker & Humblot, Berlin

European Data Watch

This section will offer descriptions as well as discussions of data sources that may be of interest to social scientists engaged in empirical research or teaching courses that include empirical investigations performed by students. The purpose is to describe the information in the data source, to give examples of questions tackled with the data and to tell how to access the data for research and teaching. We will start with data from German speaking countries that allow international comparative research. While most of the data will be at the micro level (individuals, households, or firms), more aggregate data and meta data (for regions, industries, or nations) will be included, too. Suggestions for data sources to be described in future columns (or comments on past columns) should be send to: Joachim Wagner, University of Lueneburg, Institute of Economics, Campus 4.210, 21332 Lueneburg, Germany, or e-mailed to ⟨wagner@uni-lueneburg.de⟩.

De Facto Anonymised Microdata File on Income Tax Statistics 1998

By Joachim Merz, Daniel Vorgrimler and Markus Zwick*

1. Introduction

Since microdata permit many layers of analysis, the long-standing wish of the scientific community to analyse these data in their original form as individual items of data has grown considerably over time. In particular, the confidential individual data of the income tax statistics not only are of interest for tax analyses but in addition for general income distributional analyses covering high income with higher precision than in other sources.

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

* We wish to thank the members of the FAST (*Faktische Anonymisierung der Steuerstatistik)* Advisory Board for their helpful discussion in developing the anonymisation concept.

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.

With the de facto anonymised microdata file of the income tax statistics 1998, the official statistics are expanding the standardised scientific-use files offered.

This article describes the approach followed and states reasons for the decisions leading to the de facto anonymous file of the German wage and income tax statistics 1998. Whilst Chapter 2 explains in detail the income tax statistics and the 10 % sample obtained 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[1]

Article 2 subsection 2 of the Act on Fiscal Statistics (Gesetz über Steuerstatistik – 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*)[2], in addition to the data contained in tables used to create a Federal result, the individual items of information

[1] cf. concerning the information provided in this section see Zwick, 2001, 639.

[2] 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).

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 therefore is 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 for an observation of this social group.[3]

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[4].

[3] For instance in the context of the Wealth and Poverty Report of the Federal Government; cf. on this Merz, 2001.

[4] cf. Bach / Bartholmai, 2000.

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[5] is calculated (for instance gross income minus the income-related expenses). This is currently not possible with profit income[6], 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[7].

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.[8]

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

---

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

6 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.

7 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.

8 e.g. cf. Bach / Haan / Rudolph / Steiner, 2004.

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.[9] 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.[10] 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 com-

9 Article 7 subsection 4 of the Act on Fiscal Statistics.

10 cf. Zwick, 1998, 570.

plete 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.

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.

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.

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.

With the aid of the total amount income, the data with the positive income were sub-divided in five ranges (cf. Overview 2). The "earnings" 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).

Overview 2

Sub-Division of the Anonymisation Ranges

Anonymisation range	Positive total amount income in EUR	Negative total amount income in EUR
1	0 to 64,106 (double the average total amount income)	0 to – 102,258
2	64,107 to 137,532 (99 % percentile)	–
3	137,533 to 970,202 (99.95 % percentile)	– 102,259 to – 511,292
4	970,203 to 7,354,714 (up to the 1,000 richest)	–
5	> 7,354,714	< – 511,292

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.

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.

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. 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.

Overview 3

General Anonymisation Methods

Variable(s)	Methods
Reason for assessment	Re-coding of the eight attributes in: 1 = assessed cases 2 = manual cases
Religions (in each case separated for men and women)	Re-coding of the twelve attributes in: 1 = Protestant 2 = Catholic 3 = others 4 = none
Type of assessment	Re-coding of the eight attributes in: 1 = basic table 2 = splitting table
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.
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.

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 variables were 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, 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, income from trade, income from independent personal service income from dependent personal service, income from investment of capital, income from rentals and royalties, other income, special expenses which are not expenses of a provident nature, special expenses: expenses of a provident nature, extraordinary financial expense, 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 only blurred, falsified or no longer contained in the target data at all is less valuable to a data intruder than the original information.[11]

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 3 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
	3	Yes	Yes	Yes	Dummy	No

[1)] See overview 2.

[11] cf. Lenz / Sturm / Vorgrimler, 2004, 621 – 638.

The scientists involved in the discussion of the anonymisation concept prefer to conserve the continuous variables as against the 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. Because of this the continuous variables up to and including the third range could remain unchanged in the data.

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 replaced by the average values of their respective attributes (microaggregation[12]). 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.

3.3.3 Additional Informations of the Anonymised File

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.

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.

12 On microaggregation cf. Domnigo-Ferrer / Mateo-Sanz, 2002, 189.

4. Test of Data Protection

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.[13] 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.

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. In order to be able to successfully re-identify respondents of an anonymised file, for a data intruder need additional outside knowledge of the specific respondent, knowledge of participation and variables which are contained both in external and in target data (key variables).[14]

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. For individual intrusions with specific groups of individuals, however, sufficient additional information is available. 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. Because of these arguments, the protection analysis focuses on individual intrusions on the groups “famous personalities, company executives”; “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 sub-divided into the following four categories:[15]

1. high-quality easily-accessible additional knowledge (prime keys)
2. low-quality easily-accessible additional knowledge (background keys)

---

13 More information on testing for data security of the anonymised income tax statistics 1998 cf. Scharnhorst / Zühlke / Stegenwaller, 2006.

14 Brand / Bender / Kohaut, 1999, 57 – 74.

15 cf. Elliot / Dale, 1999, 6 – 10.

3. hard-to-access high-quality additional knowledge (critical keys)
4. hard-to-access low-quality additional knowledge (inefficient keys).

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*. The *critical keys* include information on the age of the first three children, religion, donation activity and maintenance obligations.[16]

4.3 Results of Re-Identification Attempts

Famous People and Company Executives

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. 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.

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

[16] For detailed reasoning on this sub-division cf. Scharnhorst / Zühlke / Stegenwaller, 2006.

North Rhine-Westphalia. 15 of these individuals could be attributed, 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. For this reason, as additional protective measures all respondents identified as Members of Parliament have been assigned anonymisation range 5.

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. In none of the three individual intrusions on personal environment could a clear attribution be achieved. For this reason, the re-identification attempts could also be regarded as having failed.

The re-identification attempts carried out on the basis of individual intrusions have shown that the respondents are de facto anonymous. For this reason the data 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 for various economic and social research. 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* (http://www.forschungsdatenzentrum.de/). 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 Federal Statistics Law (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 pos-

sible 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.[17] 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'.

[17] cf. on this Zühlke / Zwick / Scharnhorst / Wende, 2004, 567 – 578 as well as http: // www.forschungsdatenzentrum.de.

Summary and Recommendation

By creating a scientific-use file of wage and 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.

References

Bach, S. / *Bartholmai,* B. (2000): Möglichkeiten zur Modellierung hoher Einkommen auf Grundlage der Einkommenssteuerstatistik, DIW-Diskussionspapiere No. 212.

Bach, S. / *Haan,* P. / *Rudolph,* H.-J. / *Steiner,* V. (2004): Reformkonzepte zur Einkommens- und Ertragsbesteuerung: Erhebliche Aufkommens- und Verteilungswirkungen, aber relativ geringe Effekte auf das Arbeitsangebot, DIW-Wochenbericht Vol. 16.

Brand, R. / *Bender,* S. / *Kohaut,* S. (1999): 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, 57 – 74.

Domnigo-Ferrer, J. / *Mateo-Sanz,* J. M. (2002): Practical data-oriented microaggregation for statistical disclosure control, IEE Transaction on Knowledge and Data Engineering 14 (1), 189.

Elliot, M. / *Dale,* A. (1999): Scenarios of attack: the data intruder's perspective on statistical disclosure risk, Netherlands Official Statistics, 6 – 10.

Lenz, R. / *Vorgrimler,* D. (2004): Geheimhaltungsmethoden auf dem Prüfstand – Eine Analyse anhand der Umsatzsteuerstatistik, Wirtschaft und Statistik 6, 623 – 624.

Lenz, R. / *Sturm,* R. / *Vorgrimler,* D. (2004): Maße für die faktische Anonymität von Mikrodaten, in: Wirtschaft und Statistik 6, 623 – 624.

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.

Scharnhorst, S. / *Zühlke,* S. / *Stegenwaller,* L. (2006): Beiträge zum Projekt "Faktische Anonymisierung der Lohn- und Einkommensteuerstatistik 1998", appears in the series FDZ-Arbeitspapiere, http: // www.forschungsdatenzentrum.de.

Vorgrimler, D. / *Zwick,* M. (1998): Faktische Anonymisierung der Steuerstatistik (FAST) – Lohn- und Einkommensteuer 1998 –, appears in the series FDZ-Arbeitspapier, http: // www.forschungsdatenzentrum.de.

Zühlke, S. / *Zwick,* M. / *Scharnhorst,* S. / *Wende,* T. (2004): The research data centres of the Federal Statistical Office and the statistical offices of the Länder, Schmollers Jahrbuch 124 (4), 567 – 578.

Zwick, M. (2001): Individual tax statistics data and their evaluation possibilities for the scientific community, Schmollers Jahrbuch 121 (4), 639 – 649.

Zwick, M. (1998): Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken, Wirtschaft und Statistik 7, 566.