STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE
EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**
(Skopje, The former Yugoslav Republic of Macedonia,
14-16 March 2001)

Working Paper No. 25
English only

Topic III: Attitudes of respondents towards statistical confidentiality

STATISTICAL MICRODATA – CONFIDENTIALITY PROTECTION VS FREEDOM OF INFORMATION

Invited paper

Submitted by Statistics Sweden[1]

Abstract: The paper discusses how a statistical office could strike a satisfactory balance between confidentiality protection and freedom of information. Flexible use of statistical data is of vital interest for researchers and for the democratic process. On the other hand, the willingness of respondents to provide data is dependent on the ability of the statistical office to guarantee their anonymity.

The paper argues that a combination of measures of different kinds are needed: legal, administrative, methodological, and technical. As long as statistical data are at all collected and statistical results are published, the risks of inadvertent disclosures of information about identifiable individuals (persons or enterprises) cannot be completely eliminated. On the other hand, the motivation to spend a lot of efforts to break through protection measures is usually low, especially if such efforts are regarded as criminal and can be punished. Moreover, there are often easier ways to find out sensitive information about individuals than by means of malicious processing of statistical data.

The paper presents two new ideas that are being launched and discussed in Sweden right now: (i) the idea of transforming commonly known identifiers (of persons and other objects) into pseudoidentifiers by means of a table or an algorithm that is known only by the statistical office; (ii) the idea of a statistical firewall, which filters the queries from users of statistical data as well as the statistical outputs resulting from these queries, thus monitoring the traffic between external users and internal databases containing sensitive statistical microdata. It is discussed in the paper how these two ideas can be used in practice, increasing legitimate usage and improving confidentiality protection at the same time.

## I.	THE CONFIDENTIALITY/PUBLICITY DILEMMA OF A STATISTICAL OFFICE

1.	A major problem for a statistical office is to make an optimal trade-off between, on the one hand, the confidentiality protection of the data that it collects from its respondents – the statistical microdata – and, on the other hand, the freedom that it provides to its users to get access to these data for statistical and analytical purposes.

2.	Two seemingly easy solutions to the problem would be, either (a) to remove any identifiers from the microdata before they are released, or (b) never to release any microdata at all. However, none of these solutions would be satisfactory, and even less they would be optimal.

3.	Solution (a) would not prevent an intruder from re-identifying a lot of the microdata that have been anonymized by the removal of identifiers. As has been illustrated by numerous scientific articles over the last 30 years, there is almost always a possibility for an ill-willing intruder to re-identify (part of) anonymized microdata, by means of more or less sofisticated mathematical methods. The risks are

---

[1] Prepared by Bo Sundgren.

particularly high, if the released, anonymized microdata in addition to some sensitive variables also contain some insensitive variable that are public or easily available from other sources.

4.   Solution (b) would severly restrict the usefulness of collecting data for statistical and anlytical purposes. Many powerful methods used by serious researchers and analyzers are based on the availability of microdata. Two examples are integration of data from different sources, and so-called longitudinal studies. Moreover, just replacing microdata by aggregated data, so-called macrodata, would not really be a solution either. The same scientific literature that has clearly demonstrated the possibilities of re-identifying anonymized microdata has equally clearly demonstrated the possibilities of re-identifying the microdata underlying statistical aggregates; the possibilities are particulary evident when population or sub-domains are small and/or when some variables have very skew distributions. For example, in business statistics there are often one or a few well-known companies that dominate the scene in terms of certain quantitative values. This is particularly true in small countries and in regions (states) of bigger countries. Even when the risks for re-identifying microdata behind macrodata, there may be so-called group privacy risks. Objective and correct statistical data may sometimes demonstrate facts that are embarrassing for whole groups of individuals (and thus indirectly for the individuals themselves). For example, it may actually be true that most restaurant owners are tax evaders[2], or that immigrants from a certain country have a significantly higher rate of criminality than the population at large.

5.   So the only safe way to protect the privacy and confidentiality of respondents of statistical surveys seems to be never to publish any statistical data from theses surveys at all, neither microdata, nor macrodata. Even safer would be not to collect any data from respondents in the first place – and why should we, if we are not going to publish any results? In every society there are always some people, who would welcome such solutions. For example, political extremists (to the right or to the left) find it easier to convince the general public about their ideas and prejudice, if there are no statistical facts available. A variation of this theme is that the availability of statistical facts should be restricted to those who know how to handle them in a responsible way (for example those in power in an authoritarian society). From these examples follows, on the other hand, that a well functioning democratic society requires high publicity and availability of objective, relevant, and correct statistical data. Such data form the best protection against propaganda.

6.   Unfortunately, a statistical office cannot go to the other extreme, to make all statistical data freely available. A statistical office cannot maximize the usefulness of statistical data from the point of view of its users. It must also consider the interest of the respondents. There are two aspects of the need to respect the interests of respondents by protecting the data that they have submitted to a statistical office. The first aspect is that most democratic societies put a value in itself to privacy and confidentiality; every citizen should have the right to a private sphere, and every company should have a right to some business secrets. However, for a statistical office, the second aspect is even more important. It is actually a self-interest of a statistical office to protect data submitted by respondents for statistical purposes. If a respondent does not trust the statistical office, he or she will refuse to participate in statistical surveys or provide biased data.

II.   A COMBINED TOOLS APPROACH FOR SOLVING THE DILEMMA

7.   There are at least three sets of tools that can be used for tackling the confidentiality/ publicity dilemma of a statistical office:

- Legal tools (including administrative procedures)
- Methodological tools
- Technical tools

8.   None of these sets of tools are by themselves enough to solve the dilemma in a satisfactory way. Legal tools are not enough, because sometimes it may be too easy to break the law. A recent (though non-statistical) example from Sweden is when a large number of unauthorized hospital staff members could not resist their curiosity to have a look at the journal of a well-known politician, who suddenly died at the

---

[2] Here it should be noted that in Sweden tax evasion and alcohol misuse are as sensitive topics as sex in the United States.

hospital. The misuse was detected and punished, but obviously the curiosity of the staff members was stronger than their respect for the law and fear for punishment.

9.   Methodological tools alone are not enough either. Like in espionage there is an ongoing race between "spies" and "counter-spies", that is, between those who device confidentiality protection methods and those who invent new methods of breaking the protection methods. Furthermore, the protection methods are often very complicated and not easy to understand for the general public. The general citizen just has to trust the scientist responsible for the method, and why should he or she trust a scientist in this role, if he or she is not prepared to trust a scientist in the role of user/analyst of statistical data? Moreover, some methods manipulate the statistical data in some way or another. This creates a pedagogical problem to explain why a statistical office first spends a lot of money to remove errors from the data, and then spends more money to deliberately introduce errors to make the data safe from the respondent's point of view.

10.   Finally, as we all know, technical protection methods are never perfect either.

11.   Fortunately, for the sake of statistical offices, it may be possible to achieve a high level of data protection and a high level of data availability at the same time by combining tools from the three categories mentioned above. I propose that the purpose of this trade-off between two contradictory goals is to maximize the availability of useful statistical data of good quality to a number of important processes in society within the restrictions given by legislation and the self-interest of statistical offices to preserve their confidence and good reputation with its respondents.

12.   The processes to be supported by statistical data include

- democratic processes (e.g. public debate)
- research processes
- business processes (both private and public)
- education processes
- other knowledge formation processes

13.   The rest of this paper will illustrate how such a trade-off can be achieved, taking the situation in Sweden as an example. The discussion will be structured according to the three sets of tools that were mentioned above and that can be used for improving both data protection and data availability:

- legislative tools (including administrative procedures)
- methodological tools
- technical tools

III.   LEGISLATIVE TOOLS

14.   The legislative situation is different in different countries. In Sweden there are fundamental laws and other laws. The fundamental laws may be said to form the Swedish constitution, and they can only be changed in rather complicated ways. A rather unique Swedish feature is the principle of publicity, which belongs to the fundamental laws. The meaning of this principle is, simply expressed, that all information that is created in the public sector, should be public, e.g. all decisions taken by a public authority and the documents behind the decisions, all correspondence coming into and going out from a public authority. Every public authority should have a public register to facilitate the search for correspondence and documents. Anyone can demand to see the register as well as any document without having to reveal his/her identity or why he or she wants to get the requested information.

15.   The purpose of this very powerful principle of publicity is that the citizens (often with the help of journalists) should have good possibilities to examine and question decisions taken by public authorities.

16.   All exceptions to the principle of publicity must be stated explicitly in special laws. One exception concerns data that have been submitted to an authority for statistical purposes. Such data is secret and must not be disclosed, if there is any risk that the respondents may be hurt if the data are

disclosed. In practice this means that Statistics Sweden (or any other authority in Sweden producing statistics) will not disclose statistical microdata to anyone, for whatever purpose[3], unless certain conditions are fulfilled. According to the present practice, the most important conditions are that

- there is a respectable purpose for disclosing the data, typically research performed by a university institution, but data may also be disclosed for analyses performed by other agencies and institutes, both public and private
- data have been anonymized in such a way that the individual records may not easily be re-identified
- the receiver of the data agrees to treat the data according to certain rules stated in a contract

17.   The condition that it should not be easy to re-identify data may seem to be surprisingly week. However, there is another law that criminalizes any attempt to re-identify statistical data. In combination with the other conditions stated above, this law ensures, beyond reasonable doubt, that the disclosure of anonymized data to researchers and equals will not in any way harm the respondents. Why should a researcher commit a crime and spend a lot of time and resources in order to re-identify some data received for research purposes? And if the researcher would actually have a reason to risk his or her career by committing such a crime, and to use the re-identified data in such a way that it would harm the respondent, the researcher could most probably have obtained the re-identified data in some simpler way in the first place.

18.   Moreover, it is the rule that respondents to statistical surveys are informed that the data will be used for statistical purposes such as research and statistical analysis. In cases where the microdata originally emanate from administrative sources (which is often the case for statistical microdata in Sweden), the data will in most cases (according to the principle of publicity) already be publicly available from adminstrative agencies (e.g. tax authorities). Nevertheless, once the data have arrived at Statistics Sweden, they will be treated as data submitted for statistical purposes and released only in accordance with the laws and policies stated above for such data.

19.   It can be noted here that the principle of publicity makes a lot of administrative data (e.g. income and taxation data) public in Sweden, whereas in many other countries such data are secret. For example, many newspapers in Sweden regularly publish detailed list of the income, wealth, and tax debts of famous people. This circumstance also has the effect that it would be extremely easy (in comparison with the situation in many other countries) to re-identify statistical microdata (and even a lot of aggregated statistical data for that matter) for a person with motivation and resources. From the methodological literature we know that in theory it is almost always possible derive some sensitive data concerning identified individuals from anonymized or aggregated statistical data (microdata or macrodata) under these circumstances. Thus the criminalization of attempts to re-identify statistical data is of great help when it comes to releasing anonymized statistical microdata for research and comparable purposes.

20.   Another important problem that has been solved recently in Sweden by means of legislation concerns the updating of anonymized microdata. If the identifiers are just removed from statistical microdata, without any further action, it will of course be impossible to update the anonymized microdata with, say, new time versions of certain variables. This is a problem for so-called longitudinal studies as well as for research projects where one has the need to enrich the anonymized microdata, step by step, with microdata from other sources than the original ones.

21.   The new legislation, which is part of a new, comprehensive Statistical Law, opens the possibility for the responsible statistical agency (Statistics Sweden) to replace the real identifiers (e.g. person numbers) with pseudo-identifiers (e.g. informationless integers created by a table or by encryption of the real identifiers) in connection with the anonymization of microdata, before they are released to external users. A translation key between pseudo-identifiers and real identifiers may be created and kept by the responsible statistical agency. When new microdata concerning the same individuals become available

---

[3] The law has been tested in connection with crimes, when the police has demanded access to microdata from respondents to statistical surveys in order to provide evidence to the court. Even in the case of very serious and brutal crimes, Statistics Sweden has turned down all such demands, and such decisions have never been overruled by higher courts.

from other sources or for other time periods, the responsible statistical agency may, thanks to its possession of the translation key, replace the real identifiers of these data with the same pseudo-identifiers that the users already have for the data that need to be updated.

IV.   METHODOLOGICAL TOOLS

22.   Unfortunately most methodological research work in the area of statistical confidentiality has aimed at demonstrating all realistic (and sometimes maybe not so realistic) possibilities to re-identify anonymized or aggregated statistical data. It has obviously not been equally exciting and rewarding for researchers to find methods for protecting the confidentiality of statistical data without harming the availability and usefulness of the data too much. And when researchers have proposed such methods, they have often turned out to be useless in practice.

23.   Once again, it is important to note that methodological tools alone are not likely to solve the problem tackled in this paper – to find a reasonable trade-off between confidentiality protection and data availability. If we accept that confidentiality protection methods must always be adequately supported by legislation and technical tools, we may find some promising methodological approaches today. One methodological tool that may have the potential to play such a role is the ARGUS system from Statistics Netherlands. Systems like ARGUS are not able to eliminate all risks for inadvertent disclosures from anonymized microdata. However, they are able to control the risks, provided that the statistics producer is able to provide certain critical inputs, such as the degree of sensitivity of different variables in the set of microdata under consideration. Given these inputs from the responsible producer, the software is able to propose adequate actions that are consistent with the producer's judgment.

24.   Similar thoughts like those behind the ARGUS system can be found in an early paper by Block & Olsson (1976), Statistics Sweden. They defined and used the concept of "the identifying power" of variables, later called "key resolution" by Bethlehem et al (1990).

V.   TECHNICAL TOOLS

25.   Naturally technical tools are needed for the practical implementation of legislative and methodological tools. For example, we just mentioned ARGUS as a methodological tool, but it is also a technical tool in the sense that it implements and supports a certain methodology for confidentiality protection.

26.   However, technical tools can also be used in a more creative and independent way in connection with maximizing the availability of statistical microdata within the restrictions given by laws and policies concerning the confidentiality of such data. Practically speaking, technical tools could, and should, be used to ensure that the restrictions just mentioned will "disturb" the availability of statistical microdata as little as possible.

27.   Let me take an example from a recent investigation where we interviewed a number of representative users of statistical microdata. They all accepted that Statistics Sweden must protect the confidentiality of statistical microdata by trying all requests for release of such data very carefully in accordance with the rules stated above. They also admitted that most requests for release of microdata for research and equivalent purposes that they submitted to Statistics Sweden finally resulted that they received a set of anonymized microdata – not necessarily though the microdata that they had envisaged. The data had typically been made less detailed (e.g. by consolidating different values of a variable into one range of values). It was not always obvious for the users how Statistics Sweden had come to its decision, and they also suspected that different decisions concerning different kinds of statistical data or different users and usages were not always consistent with one another. Furthermore they complained that the administrative procedures behind the decisions often took an unacceptable long time and involved lengthy discussions both with the users and between different organizational units within Statistics Sweden.

28.   The criticism just described indicates a lack of objectivity and transparency in the decision process concerning release of statistical microdata from Statistics Sweden. We found that criticism

justified, and as a possible remedy we proposed an in-depth analysis that should hopefully result in a more formalized, better documented, and thus more transparent decision process. One way of achieving this in practice would be to develop some computer-support to the decision process. It may not be possible to automate the whole process, but computer-support would make the process more streamlined and faster. Moreover, it would be easier for the users to foresee the results of a more formalized process. It might even be possible to provide an expert system to the users, helping them to prognosticate the likely outcome of their request.

29.   One way to formalize the process is to identifiy the background variables, i.e. the variables which could be used for re-identification. The distributions of these variables are well known and well tabulated, so the key resolution for each of them could easily be calculated in advance. The effects of rounding or perturbation are also easily calculated. Then the key resolution of the combination of all background variables could be calculated, and the number of unique individuals in the population, and thus the risk of re-identification could be estimated. These estimates could then be used in order to facilitate the decision whether to release certain anonymized microdata or not.

30.   Another improvement of the management of statistical confidentiality, much wanted by the users according to our interviews, would be to try "kinds of requests" rather than very specific requests for certain statistical microdata to be used for very specific purposes. Such more general permission for a user to use a more broadly defined set of microdata for a more broadly defined purpose would be of great help for researcher, who often change their hypotheses and research focuses as a result of the research process itself; one analysis leads to new questions, etc, in a never-ending interactive process between the researcher and the data. Such broader release decision may require the responsible statistical office to develop and install some kind of software that would dynamically monitor the usage of the statistical microdata concerned and make sure that all requests from the users are within certain acceptable limits as concerns disclosure risks etc.[4]

31.   Figure 1 visualizes how a more general and more automated monitoring of the confidentiality and availability of statistical data could be implemented at a statistical office like Statistics Sweden. A fundamental software component is the so-called statistical firewall. It will ensure that only anonymized microdata and macrodata[5] that satisfy disclosure risk criteria set by the responsible producer will be released to external users.

32.   If we walk through figure 1 from left to right, we come first to the two major sources of statistical microdata, administrative data sources (indirect sources) and statistical surveys (direct sources). In Sweden administrative sources account for more than 95% of the statistical microdata used for official statistics production.

33.   The identified microdata stored in the databases of the statistical office may be divided into primary registers and secondary registers. Primary registers contain data that come from the sources after some preparation processes: data entry, coding, editing etc. Secondary registers contain data that are derived from primary registers by more or less sophisticated processes. A typical secondary register may contain data obtained by record linkage of data concerning different variables of the same objects in different register or different time versions of the same variables of the same objects. Secondary registers have become very popular among researchers, but they require a lot of quality-related work. For example, when records from different registers are linked to each other, inconsistencies may be detected that have to be reconciled, and when different time versions of "the same" data are put together, coding errors may give a completely false impression of the dynamics in the population.

---

[4] Cf systems like ARGUS where the disclosure risks are considered once and for all, that is, in a statical way.
[5] As has already been pointed out, the problems of statistical confidentiality are not limited to microdata.
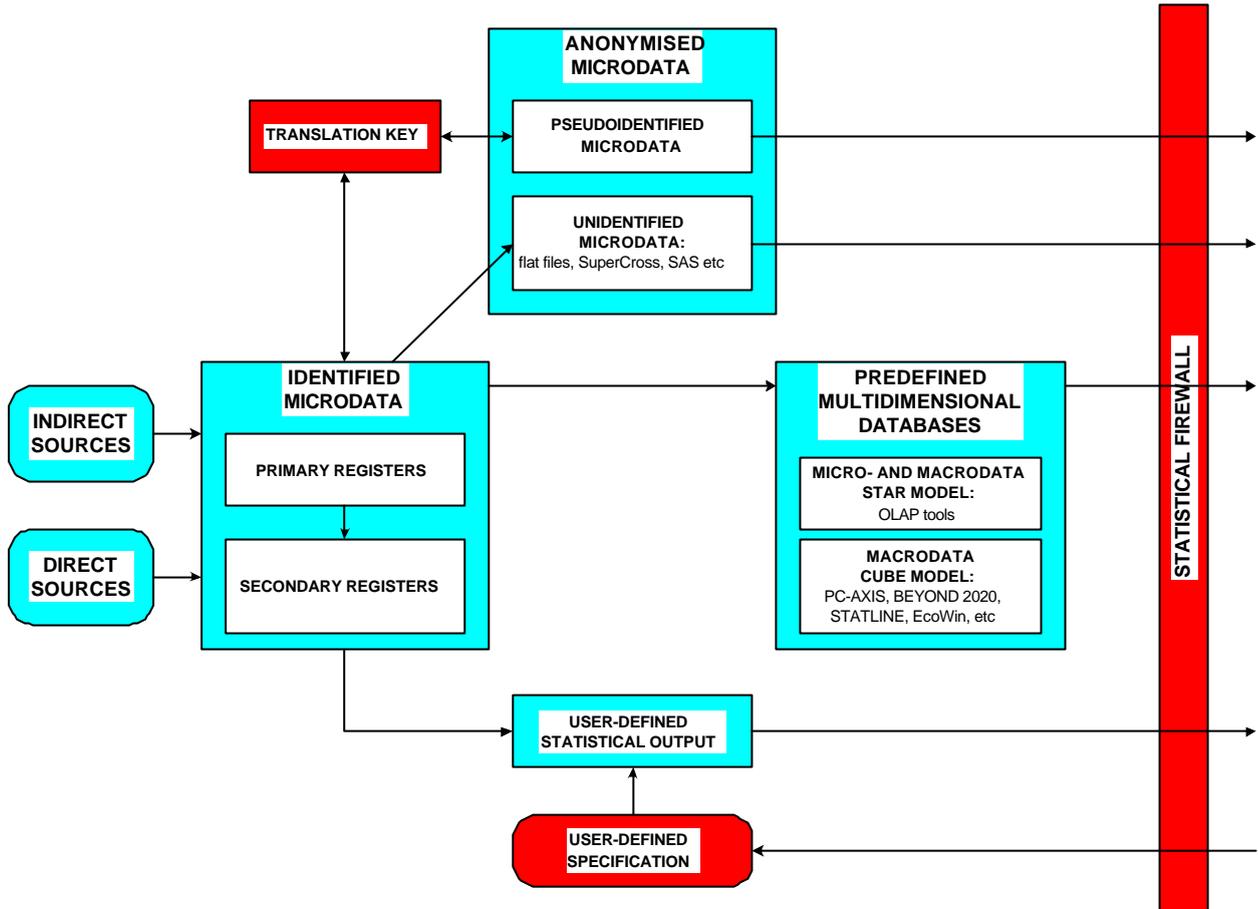
Figure 1. *Access to more or less processed statistical microdata via a statistical firewall.*

34.   Identified statistical microdata may be transformed into two kinds of anonymized microdata sets: microdata with and without pseudo-identifiers. Both types of microdata set may be (statically) checked for disclosure risks by means of tools like ARGUS, before they are released to users.

35.   Another way of controlling data confidentiality, while preserving a lot of usage flexibility, is to create predefined, multidimensional databases, either with a combination of microdata and macrodata according to the star model, managed by so-called OLAP tools, or with aggregated data only according to the cube model, managed by tools like PC-AXIS, Beyond 2020, StatLine, EcoWin, etc.

36.   The highest level of usage flexibility is achieved if the user is allowed to specify any output requests that are logically possible to formulate, given the contents of all microdata in the data warehouse. This level of flexibility also requires the most advanced level of (dynamical) confidentiality control.
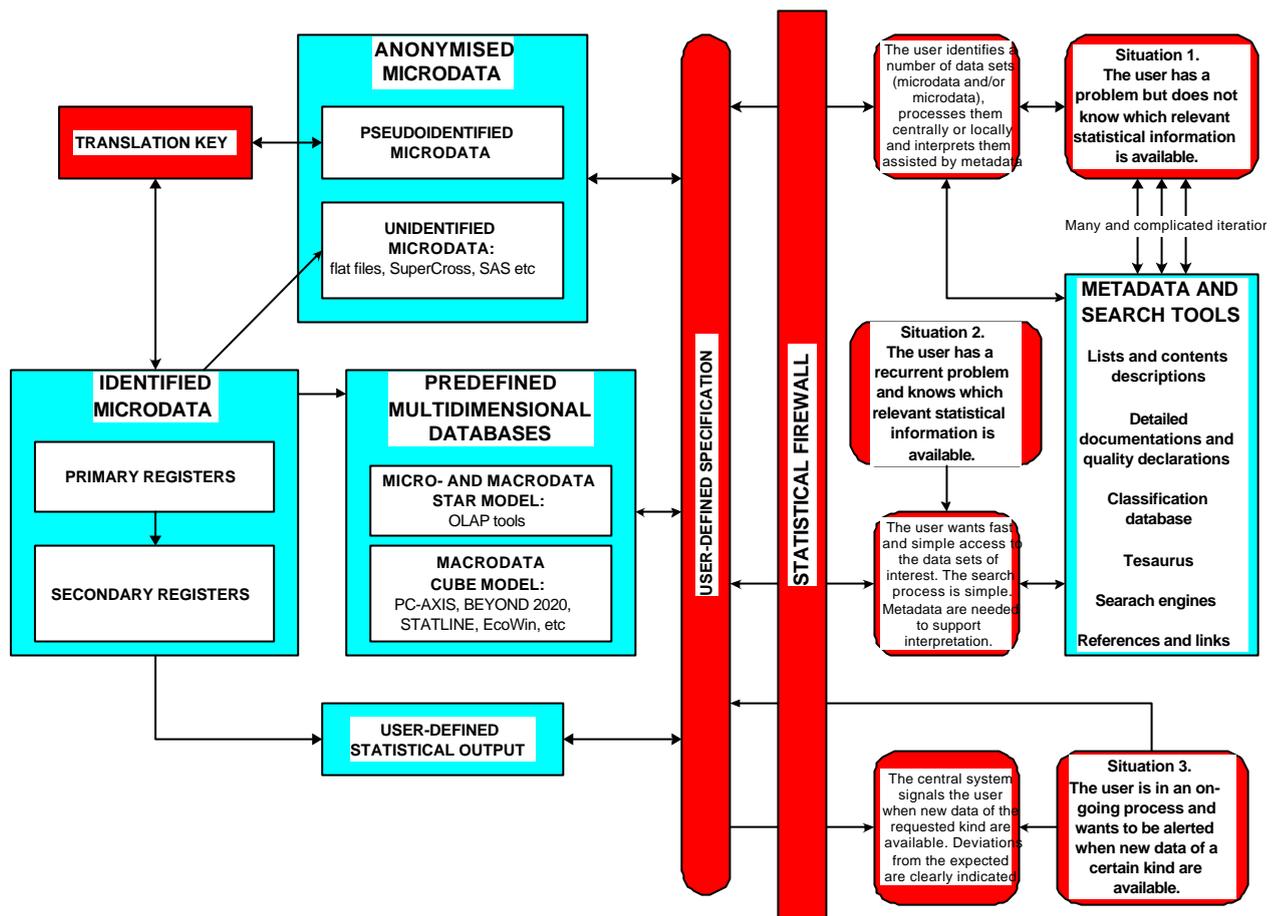
Figure 2. *The statistical data warehouse from a user's point of view. Three typical usage situations are illustrated.*

37.   Figure 1 illustrates a statistical data warehouse from a producer's perspective. Figure 2 visualizes the same warehouse of raw and processed statistical microdata from a user's point of view. More precisely three typical use situations are shown schematically:

   (i)   The user has a unique problem but does not know what relevant statistical data are available. Example: a research project or a unique problem solving or evaluation situation.
  (ii)   The user has a recurrent problem and knows what relevant statistical data are available. Example: a regularly performed analysis of the economic climate in a country on the basis of certain pre-defined indicators.
 (iii)   The user is part of an on-going business process and wants to be made aware of new data as soon as they occur. Example: a trader on the stock market.

VI.   CONCLUSION

38.   By combining appropriate legal, administrative, methodologic al, and technical measure, a statistical office can *both* increase the the availability of microdata for a wide range of statistical usages, including support for the democratic processes, *and* improve the confidentiality protection to the benefit of the respondents.

References

Bethlehem, J. G., Keller, W. J., Pannekoek, J. (1990): *"Disclosure control of microdata"*, Journal of the American Statistical Association, Vol. 85, pp. 38-45.

Block, Hans & Olsson, Lars (1976): *"Bakvägsidentifiering" / "Backwards identification"* (in Swedish with summary in English), Statistisk Tidskrift / Statistical Review, 1976, pp. 135-144.

Block, Hans (2000): *"Reidentification of perturbed microdata files"*, Paper prepared for the JSM 2000, Indianapolis, USA.

Johansson, Sten (1990): *"Information needs for the market and for democracy."* Journal of Official Statistics, Vol 6, No 1, 1990, pp 89-99.

Lievesley, Denise (1999): *"Sharing and preserving data for the social sciences."* Paper prepared for the Workshop on Infrastructure needs for social sciences, held in Ottawa, Canada 1999.

OECD Secretariat (1999): *"Social sciences databases in OECD countries – an overview."* Paper prepared by Jun Oba, consultant, for the Workshop on Infrastructure needs for social sciences, held in Ottawa, Canada 1999.

Sundgren, Bo (1999): *"Increasing the availability of Sweden's official statistics"* (Available only in Swedish: *"Ökad tillgänglighet till Sveriges officiella statistik."*) Appendix to the evaluation of the reform of Sweden's official statistics, SOU 1999:96.

Sundgren, Bo (1997): *"Sweden´s Statistical Databases – an infrastructure for flexible dissemination of statistics."* Report to the UN/ECE Conference of European Statisticians, Geneva 1997.

Wagner, Gert G. (1999): *"Producing statistical data by means of competition within a publicly funded and controlled system of statistical infrastructure – an outline of a vision."* Paper prepared for the Workshop on Infrastructure needs for social sciences, held in Ottawa, Canada 1999.