Dissemination of Official Statistics as Open Government Data Strategy

Lohauß, Peter
State Statistical Institute Berlin-Brandenburg
Alt-Friedrichsfelde 60
Berlin, 10315, Germany
E-mail: peter.lohauss@statistik-bbb.de

Official statistics in the European Statistical System (ESS) must be disseminated in a timely and punctual manner and should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance. These principles of the European Statistics Code of Practice set the frame, but define not clearly what and how to disseminate. According to some well known definitions of Open Government Data Principles[1] the dissemination process shall be more thoroughly specified.

Complete: All public data are made available. Public data are data that are not subject to valid privacy, security or privilege limitations.

Official Statistics is the provider of a large amount of official data. The first question is how many data of a given survey are made publicly available? Usually only the most common variables and items are to be published as tables. There were several good reasons not to disseminate the complete body of data. First, statistical confidentiality does not allow publishing single observations (or numbers smaller than 3), that means, the microdata must be kept secret. Second, many users demand understandable tables, but not a bunch of hardly readable figures. Third, the huge amount of all existent figures makes it impossible to publish it on paper. But things have changed. With the emerging of a new class of users of official statistics - experts and scientists - and the development of convenient IT-tools for statistical analysis there is a strong demand to get access to microdata and there is enough knowledge and intelligence on the side of many users to deal with huge amounts of microdata. Some users can make even more use of Official Statistics than the NSAs, because they mobilize more expertise and manpower than the Offices themselves. They can make the whole body of figures transparent to the public and are able to enhance the benefit and value of Official Statistics. To match the criteria of complete and public data Official Statistics need solutions for a safe access to confidential microdata.

Primary: Data are collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.

The second question is: what is the source of statistical data or what are raw data? To specify this, a look at the steps of the generic statistical business process model[2] will help. After survey data are reviewed, validated and edited several sub processes have to be done to transform these raw data into statistically

useful data. Missing and wrong values are imputed, new variables and statistical units are derived, weights are calculated to gross up sample survey results to make them representative of the target population, aggregates are calculated, summing data for records sharing certain characteristics, determining measures of average and dispersion, and applying weights to sample survey data to derive population totals. The data files are finalized to obtain the input for analysis. When these steps are done one more sub-process is necessary to finalize the output: it ensures the statistics and associated information are fit for purpose and reach the required quality level, and are thus ready for use. It includes completing consistency checks; determining the level of release, and applying caveats, measures of uncertainty, and all necessary metadata.

The source of statistical data is not the primary state after editing, but includes all necessary steps to finalize the data as statistical results. The degree to which the raw data are transformed can be documented in the metadata.

The finest possible level of granularity is microdata. To observe statistical confidentiality requires methods of safe access or aggregation to levels not containing confidential data. Since all aggregation and applying methods of confidentiality diminishes the level of information and use they have to be applied carefully.

Accessible: Data are available to the widest range of users for the widest range of purposes.

Accessibility of public data is to be granted online and via the internet. It is managed by implementing and updating of systems which store data and metadata for dissemination purposes, this includes formatting data and metadata ready to be put into output databases; loading data and metadata into output databases and ensuring that data are linked to the relevant metadata. Of all types of access (publication, consulting, on-site access, etc.) online access is the one which serves the widest range of users.

Machine processable: Data are reasonably structured to allow automated processing.

Official Statistics can be presented as downloadable tables in data files in common formats or as web-based databases with output to be downloaded in common formats. The distribution of microdata files is severely restricted due to statistical confidentiality. Solving the task of giving users access to microdata has long deemed impossible, but ways of controlling the confidentiality are developing. Most Regional Data in Official Statistic can be defined as geo specialist data according to the Annexes of the INSPIRE direction of the EU. These Data should be collected only once and kept where they can be maintained most effectively. It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications. This requires the development of databases with the content of Official Statistics that provide machine processable geo specialist data.

There are some more principles of Open Government Data which are easily to fulfil because they are part of the Quality Criteria of Official Statistics: This applies to the principles of *timeliness*, *non-discriminatory distribution* and *non-proprietary formats*. The last principle *license-free* is met

since the greater part of Official Statistics Data is distributed free. Official Statistics is a common good and the use of official statistical data should be free for the public.

A strategy to meet the demands of open government principles in the dissemination of Official Statistic will comprise different key elements. Up to date Official Statistic in Germany has achieved a set of approaches. They comprise different ways and levels of anonymisation and information for use. The Statistical State Institute Berlin-Brandenburg (SSI Berlin-Brandenburg) offers these four main ways of distribution:

- **Presentation of online tables in downloadable formats**

Statistics online is offered for download as xls-files or pdf. By changing the content of our web-site from html-tables to downloadable formats we provide the public users with a content which is fit for further use.

- **Implementation of online data bases providing flexible queries and analysis**

Flexible analysis for expert users is provided as an online database. SuperWEB™ is a Web solution for ad hoc tabulation, providing access and analysis of large privacy protected, confidentialized microdata tables. SuperWEB is designed for people who want to ask their own questions and create their own reports rather than rely on information produced by other people. SuperWEB provides access to all the available data and allows users to follow a Query-Answer-Query process. SSI Berlin-Brandenburg, as a provider of regional statistics for the city of Berlin, gives expert users SuperWEB access to the cities population register as a micro-data confidentialized data base[3].

- **Distribution of microdata files as Public Use Files (PUF) or Scientific Use Files (SUF)**

As absolutely anonymised microdata, standardised Public Use Files (PUF) are available to all those who are interested both within the country and abroad. Due to anonymisation, Public Use Files contain only selected variables. As a rule, variables with a high degree of subject-related detail are aggregated. In most cases, Public Use Files do not allow detailed regional breakdowns.

The research data centres offer standardised Scientific Use Files (SUF) as microdata of common statistics which are de facto anonymised for off-site use to users from the scientific community. These data have a far greater information potential than Public Use Files and they are well-suited for the large part of scientific data analyses. Off-site use is possible at research institutions which are governed by German law. Foreign data users who are not employed by a German research institution may work with de facto anonymised SUFs via remote execution or at the safe centres in the statistical offices.

- **Safe Centres and Remote execution in Research Data Centres**

De facto anonymised microdata can be analysed by domestic and foreign guest scientists on the protected premises of the statistical offices. De facto anonymity is achieved not only by an anonymisation of the data but in combination with a controlled data access to give much more detailed information than the Scientific Use Files.

Remote execution is the only way of access permitting the analysis of formally anonymised original data. However, the data user does not have direct access to the data. The data users receive structural data records (dummy files) which resemble the original material with regard to structure and the values of the variables. With the help of these dummy files, evaluation programs (syntax scripts) can be prepared using the analysis programs SPSS, SAS or STATA, which will then be used by the statistical offices to analyse the original data. After the required confidentiality checks have been made, the data users finally receive the results of that analysis. By means of remote execution, all interested parties (both within the country and abroad) may analyse microdata of official statistics.

Further development comprises research about novel approaches in providing remote access to micro data in official statistic. The MORPHEUS system[4] is an anonymisation technique that will provide access to micro data and returning results in real time. This would allow changing the remote access to online access.

REFERENCES (RÉFERENCES)

[1]http://opendata-network.org/2009/11/open-government-data-principles/.

[2] Generic statistical business process model (GSBPM) version 4 (2009); http://www.statcan.gc.ca/concepts/gsbpm-msgpo-eng.htm

[3] Anonymisation of microdata with SAFE: J. Höhne: SAFE A Method for Statistical Disclosure Limitation of Microdata, Working Paper No 37 Joint ECE/Eurostat work session on statistical data confidentiality

[4] Julia Höninger: Morpheus – An Innovative Approach to Remote Data Access. Working Paper at the 58th ISI World Statistics Congress