DDI Data Description Statistics Protection Software

Sebastian Kočar

Social Science Data Archives, University of Ljubljana, Ljubljana, Slovenia
sebastian.kocar@fdv.uni-lj.si

Abstract. The article presents the background, the purpose in developing the DDI Data Description Statistics Protection Software and its features, which enable easy anonymization processes of DDI (Data Documentation Initiative) standard univariate statistics. To promote detailed official statistics microdata use for scientific purposes, variable information could be publically distributed. However, since the descriptive statistics is accessible to the public, the aggregated data should be protected. Statistical software for automatic data protection was developed to address this issue. It has been programmed to perform data protection methods such as bracketing, top- and bottom-coding, variable (information) removal, numeric descriptive statistics protection and low frequency protection following the minimum frequency rule. In contrast to existing microdata and aggregated data anonymization tools, the developed software protects aggregated data directly in the XML code. One of the challenges of the future would be to programme the presented tool as an add-on for Nesstar Publisher software.

Keywords: DDI Data Description; XML; data anonymization methods; data protection software; official statistics data

1 Introduction

There is a growing interest in accessing official statistics microdata for research and study purposes. One of the projects that focus on improving the European scientific research environment by connecting registered researchers of academic/research institutions to national statistics institutes is Data without Boundaries [3]. The European Commission Framework Programme 7 funded project which brought together partner organizations from three distinctive backgrounds: social science data archives, national statistics institutes including Eurostat and researchers. Slovenia is represented by two partner organizations: Social Science Data Archives of the University of Ljubljana and the Statistical Office of the Republic of Slovenia (SORS), which are actively involved in two distinctive work packages[1]. At the same time, to improve the research environment on the national level, we simultaneously engaged the coopera-

[1] DwB WP5 - Servicing European Researchers in the use of Official Statistics Microdata (ADP) and DwB WP3 - Enhancing legal, information security and researcher accreditation frameworks for access to data (SORS); both organizations support tasks of other WPs, such as WP2.

tion of both organizations in the microdata access field, which is a continuation of our collaboration from the past[2], but has become more intense and goal-oriented since the start of the DwB project [13].

Social Science Data Archives (ADP) have the skills and experience to add value to the official statistics microdata by preparing quality metadata using the DDI (Data Documentation Initiative) standard [2] and distributing supporting documentation which makes using microdata a much easier task. ADP employees work directly with official statistics microdata in the SORS data lab, preparing data that is easy and simple to use for scientific and academic purposes. Furthermore, both organizations try to promote the use of the official statistics microdata by distributing metadata without the detailed microdata. That has been accomplished by publishing useful information on surveys and data on the ADP website, and also by using the Nesstar online tool. It is believed that providing detailed information about microdata, which could later be accessed in the research data lab at the SORS or by remote access, will encourage researchers to apply for access to microdata. The more detailed the metadata are, the better promotional value they have. Therefore, the complete DDI 2 metadata scheme regularly used in the CESSDA archives for preparing academic research metadata has also been applied for the official statistics data. That metadata scheme consists of all major DDI 2 sections: Document Description, Study Description, File Description, Other Materials and Data Description [2]. The latter is the only section which cannot be prepared without accessing detailed microdata; therefore, it gives researchers a broad overview of what is exactly available in the database even before being granted the access to detailed microdata in the protected environment.

However, the current trends in the microdata access field are of a restrictive nature. Official statistics providers have to follow a set of very strict rules and act in accordance to rigid regulations. Consequently microdata anonymization and tabular data protection play a growingly important role in the distribution of data and results of statistical analysis. While working on providing access to aggregated official statistics data in the form of DDI Data Descriptions, the need for protecting publically distributed descriptive statistics was recognized. And to perform the process in as little time as possible, and to allow people with no previous XML editing experience to execute the DDI Data Description anonymization, an automatic tool had to be developed.

2 Protection of microdata and aggregated data

Microdata, aggregated or tabular official statistics data distributed publically need to be examined and processed to respect the individual's confidentiality. That can only be ensured if the data of any kind are well protected and consequently the disclosure risk is minimized [15].

Microdata anonymization has been a fast growing science in recent years, mostly in response to the demand for distribution of confidential data. There are numerous anonymization software packages (μ-ARGUS [6], Cornell Anonymization Toolkit

[2] The organizations collaborated since the establishment of ADP in the late nineties. ADP has mostly been distributing SORS microdata and metadata.

[1], R! package sdcMicro [14], etc.) which use various anonymization methods and techniques, such as:

- Data reduction (removing variables or records, global recoding, top and bottom coding, local suppression, etc.)
- Data perturbation (micro-aggregation, collapsing or combining variables, data swapping, post-randomization, adding noise, etc.)
- Generating synthetic data (alternative approach to data protection). [7,8]

On the other hand, aggregated data protection deals with different kinds of issues, but is based on similar approaches to data anonymization, such as recoding (hierarchical recoding, non-hierarchical recoding) and/or (secondary) suppression (modular, rounding etc.). [5] Generally speaking, in the case of tabular data the protection has to follow the minimum frequency rule, p% rule and dominance rule. [4]

Several microdata anonymization and tabular data protection methods and techniques were taken into consideration in order to develop a tool that would sufficiently protect Data Description aggregated data in all the required ways. In the end we decided to use a combination of them. The selected methods and techniques are meant to solve all the confidentiality problems that we have come across when protecting Data Description statistics for final distribution on the ADP website[3]. The tool is specifically designed to work with DDI 2 XML files, which are presented and described in the next sections. In contrast to the existing data anonymization tools, DDI Data Description Statistics Protection Software does not protect either microdata or tabular data, but is specially programmed to protect single variable aggregated statistics (either categorical variable value frequencies or numeric variable descriptive statistics) by making the necessary changes directly in the XML code of the DDI Data Description.

3 DDI 2 Data Description sensitive variable information

DDI 2 Data Description is a metadata standard section which provides data users with variable information; the most important and commonly used information is listed below:

- Variable group
- Variable name and variable label
- Categorical variable values with labels and frequencies
- Numeric variable univariate statistics (min, max, mean, mode etc.) [2]

Variable group field is the only Data Description field which by default consists of manually added information on variables. It is based on the expert knowledge of the person who performs grouping of variables based on what they measure. Therefore, no protection of that field would be needed. Other fields are automatically filled in by Nesstar Publisher software [9]. Variable name gives users information about how a

[3] www.adp.fdv.uni-lj.si/en/

certain variable is named in the original official statistics database and variable label describes what a certain variable measures. The protection of those two fields is needed only in the most urgent cases, which will be briefly mentioned in the next section of this paper. Furthermore, the variable descriptive statistics are displayed based on the type of a variable: numeric variable univariate statistics are displayed for numeric variables and frequencies are displayed for nominal or ordinal (categorical) variable values. The listed statistics are the ones which could provide sensitive information and therefore need to be protected in certain cases.

4 Protection of DDI Data Description Statistics

In this section we present features of the software for automatic Data Description statistics protection, which are based both on microdata anonymization and tabular data protection methods and techniques, as well as on the specifics of official statistics confidentiality issues [15]. The following chapters describe individual software sections (tabs). They were thematically organized by the type of an anonymization techniques applied to the Data Description statistics.

4.1 Removing variable information

About the protection. Variable removal is the most extreme way of data anonymization. As it removes all the information about a variable, even the information about the existence of a variable in the original detailed database, it should be used only in the most sensitive cases. E.g. if it could misled the public into thinking that national official statistics do not protect confidentiality of individuals participating in a survey (because of the existence of a very sensible variable in an accessible database), which could lead into distrust in official statistics and consequently into refusals to participate in future non-obligatory research. A good example of a sensitive variable is *monthly net income*, especially if the information is not provided directly by the respondent, but is a result of merging survey and administrative record databases.

Besides the removal of all variable information, the tool also performs removal of numeric variable information only. It could be performed e.g. in case numeric values could represent a misleading piece of information. Statistical offices might be concerned about how public with little statistical knowledge would understand using various weighting variables (with a wide range of variable values) and their effect on the published results. Therefore it might be perceived that the detailed information on those kinds of variables should only be available to microdata users in a protected environment.

Software execution. In the case of the variable removal, the tool removes all the information from the XML code, i.e. everything written in the `<var></var>` field. On the other hand, removal of all numeric variable information leaves the `name=""` attribute and `<labl></labl>` field intact.

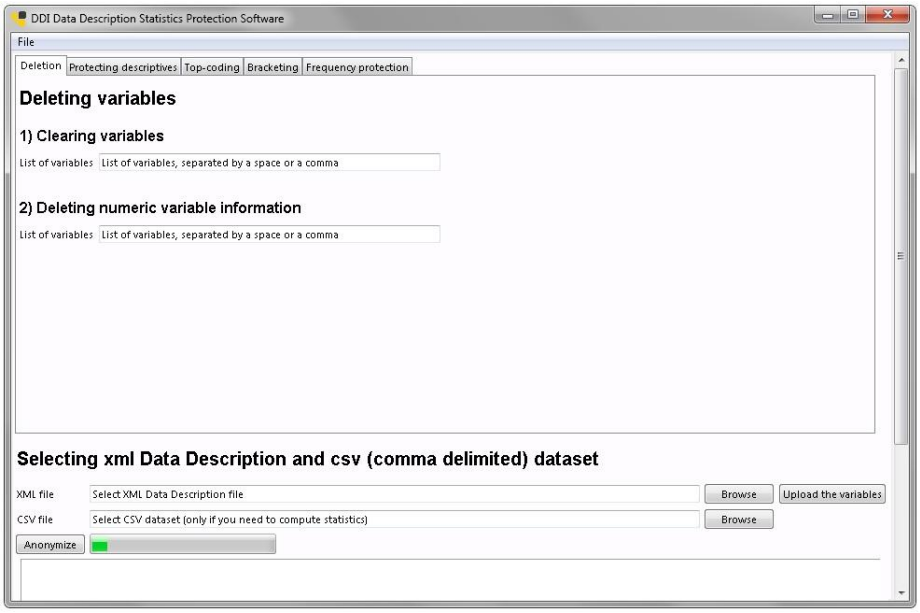

Fig. 1. Deletion tab

4.2 Protection of summary statistics

About the protection. Summary statistics are displayed for numeric variables only. The statistics calculated and added by Nesstar Publisher by default are:

- Minimum (min),
- Maximum (max),
- Mean (mean),
- Standard deviation (stdev).

In addition, DDI 2 standard enables two other statistics to be displayed: median and mode. The reasons why numeric variable information should not be displayed vary. The lowest and the highest value should be protected not to allow the identification of an "extreme" individual; e.g. if the company with far the highest number of employees in the country is included in the sample. Therefore the tool either allows protection of values by not displaying them or by replacing them with percentiles. Mean and standard deviation could, on the other hand, be protected not to misled the public; e.g. not to publish non-equal statistics as a result of analyzing weighted data for official statistics publications and non-weighted data for DDI Data Description survey metadata[4]. Therefore the software either executes protection of values by not displaying them and calculating median and mode statistics instead. The tool calculates the

[4] It is a well accepted practice in (social science) data archiving to publish original non-weighted univariate statistics.

statistics by using NumPy [10] and SciPy [12] Python [11] libraries. Users only have to upload the CSV database, which was previously used to prepare the Data Description XML file, and the libraries automatically perform the rest of the work.

Software execution. The software is programmed to replace statistics values calculated by Nesstar Publisher with "`Protected value`" text for:

```
<sumStat type="min"></sumStat>
<sumStat type="max"></sumStat>
<sumStat type="mean"></sumStat>
<sumStat type="stdev"></sumStat>
```

Additionally, it allows calculation of selected percentiles (from 1 to 99) for minimum and maximum, adding the following content to the fields (an example):

```
<sumStat type="min">6.5 (5. percentile)</sumStat>
<sumStat type="max">564 (95. percentile)</sumStat>
```

For newly calculated median and mode (as it is not possible to add them in Nesstar Publisher), programme adds the fields in the XML code and automatically calculates selected statistics. The tool uses formulas for calculating modes and medians which are included in listed Python libraries [10, 12]. The fields added are listed below:

```
<sumStat type="mode"></sumStat>
<sumStat type="medn"></sumStat>
```

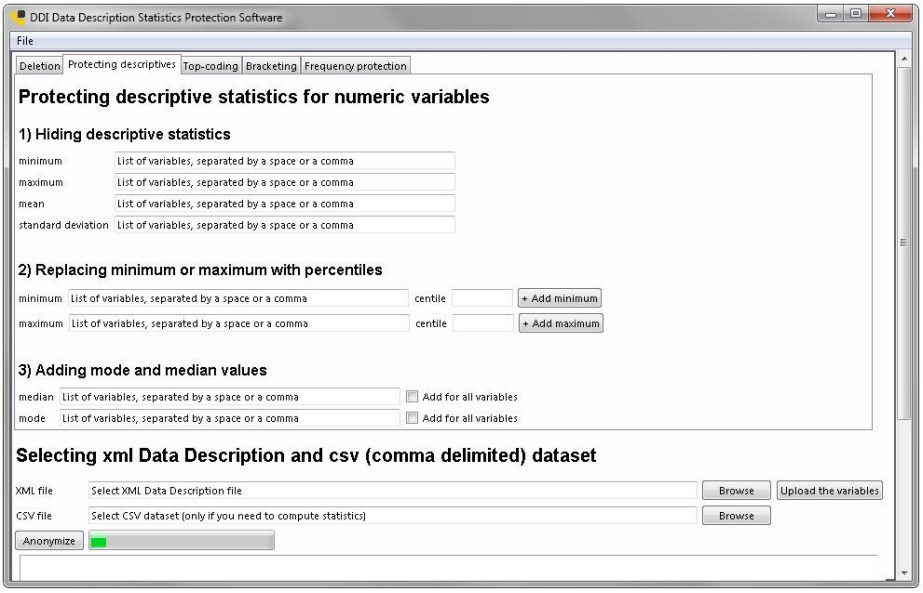

Fig. 2. Protecting descriptives tab

4.3 Top-coding and bottom-coding

About the protection. Top- and bottom-coding are performed to protect the frequencies of the highest or lowest categorical variable values. In contrast to microdata anonymization top- and low-coding techniques, numeric Data Description variables cannot and do not need to be protected that way since no variable values are categorically displayed. Top-coding and bottom-coding are in our case somewhat a less restrictive way of protecting data compared to not displaying all highest values frequencies; this way at least some information is provided. It could be e.g. used for protecting low frequencies for the highest age group values, especially in the case of relatively small survey samples, as there are not many 90+ years old respondents participating in official statistics surveys.

Software execution. The tool automatically cumulates frequencies for all values above (top-coding) or below (bottom-coding) the selected and later highest/lowest displayed value. Symbols "+" or "-" are added to the top or the bottom value. The protected code is presented below (an example):

```
<catgry>
<catValu>19+</catValu>
<catStat type="freq">12</catStat>
</catgry>
```

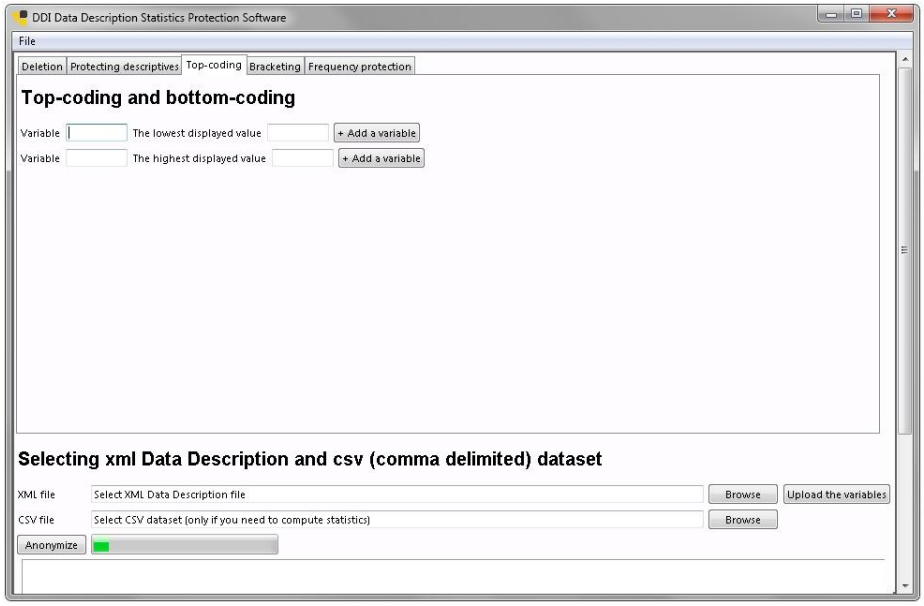

Fig. 3. Top-coding tab

4.4 Bracketing

About the protection. Bracketing, also known as recoding or collapsing variable values, is similarly to top- or bottom-coding used to protect data by combining (cumulating) frequencies of several variable values. It is commonly known that higher the frequency, better the data are protected. It is a fact that bracketing could be previously done by any microdata anonymization tool, as well as Nesstar Publisher. However, by including bracketing in the DDI Data Description Statistics Protection Software, we gave users a chance to protect aggregated data without changing the microdata matrix (which could later by uploaded into Nesstar online tool in the non-anonymized form with e.g. more strict conditions for access). We also integrated the function into the software to ensure that data protection could be done at any step, even after the final version of microdata is created. Furthermore, we believe that the whole anonymization process not only could, but also has to be done in one and only step. Consequently all the changes to the XML could be presented in one well written report, which would be automatically created by the software after the data anonymization is executed.

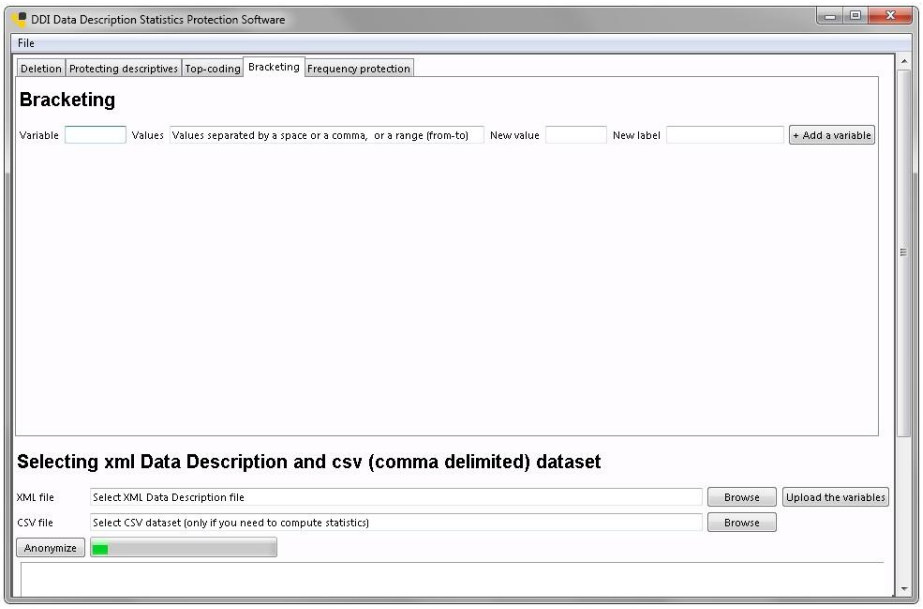

Fig. 4. Bracketing tab

Software execution. The tool automatically cumulates frequencies for recoded values. It also allows users to add labels to newly created variable values while the old value labels are automatically overwritten. The changes in the XML code are presented below.

Original code before bracketing

```
<catgry>
<catValu>1</catValu>
<labl>Bachelor's</labl>
<catStat type="freq">
400</catStat>
</catgry>

<catgry>
<catValu>2</catValu>
<labl>Master's</labl>
<catStat type="freq">
100</catStat>
</catgry>

<catgry>
<catValu>3</catValu>
<labl>PhD</labl>
<catStat type="freq">
50</catStat>
</catgry>
```

Protected code after bracketing

```
<catgry>
<catValu>1</catValu>
<labl>Bachelor's</labl>
<catStat type="freq">
400</catStat>
</catgry>

<catgry>
<catValu>2</catValu>
<labl>Master's & PhD</labl>
<catStat type="freq">
150</catStat>
</catgry>
```

4.5 Protection of category level statistics

About the protection. This function is just like the protection of tabular data based on the minimum frequency rule and plays a central role in the DDI Data Description anonymization. Users have several options on how to manage frequency display of categorical variables and the software enables the following ways of data anonymization:

- Protection of all frequencies of selected variable values (no matter the frequency displayed),
- Protection of all frequencies of all variable values (following the minimum frequency rule),
- Protection of frequencies of selected variable values (following the minimum frequency rule).

The software was programmed to provide users with several additional options which could be selected; all dataset variables could be protected the same way (using a general defined minimum frequency) or, on the other hand, groups of protected variables could be created based on various minimum frequencies thresholds defined – that is how more sensitive variables could be more restrictively protected than less sensitive ones.

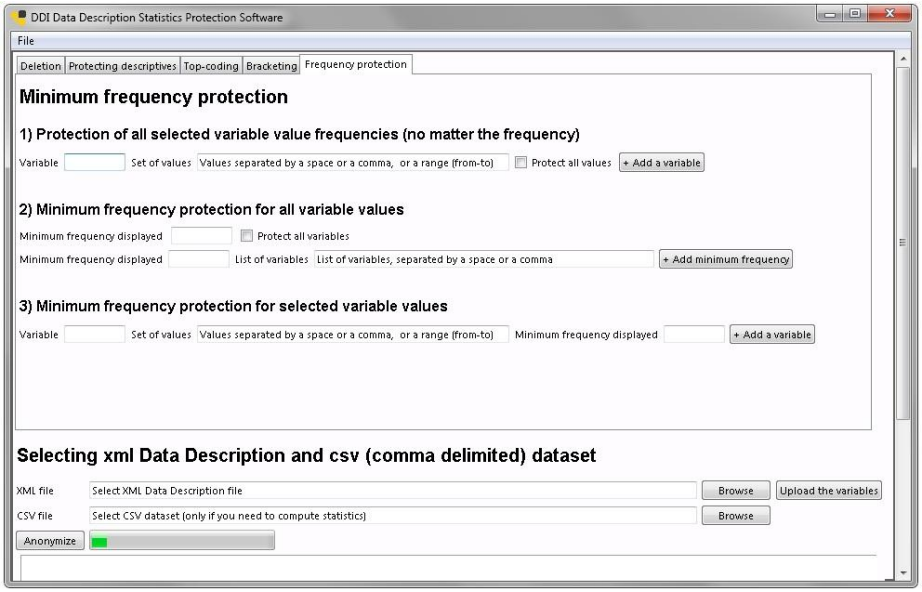

Fig. 5. Frequency protection tab

Software execution. The tool automatically finds all frequencies of categorical values which are lower than the minimum frequency still displayed (for defined variables and variable values which should be protected). Frequencies lower that the defined threshold are not displayed and are replaced by the `"Protected value"` text in the following field:

```
<catStat type="freq"></catStat>
```

It should be noted that the software had to be programmed to assure that protected frequencies could not be logically calculated from the other non-protected frequencies. Therefore the second lowest frequency is protected as well, no matter if it is higher than the threshold value. In case the sum of two lowest protected frequencies is still lower than the selected minimum frequency, additional next lowest frequencies are protected until the sum reaches the threshold.

4.6 Other features of DDI Data Description Statistics Protection Software

The most notable function added to the software is the display of all variable descriptive statistics. For that purpose, an XML Data Description file has to be uploaded into the system. When the file is uploaded, a new tab is created. It includes a frame with a list of all variables in the Data Description file and a frame with variable statistics displayed, either numeric or categorical ones. The tab is designed to show statistics of one variable at a time. The main idea of adding this function was to make the process of data protection as quick and simple as possible. The added function ena-

bles users to browse through variable statistics, to identify sensible information and to manually add variable(s) to one of more data protection tabs. This function could as well be used to double-check if the final version of Data Description XML file is properly anonymized.

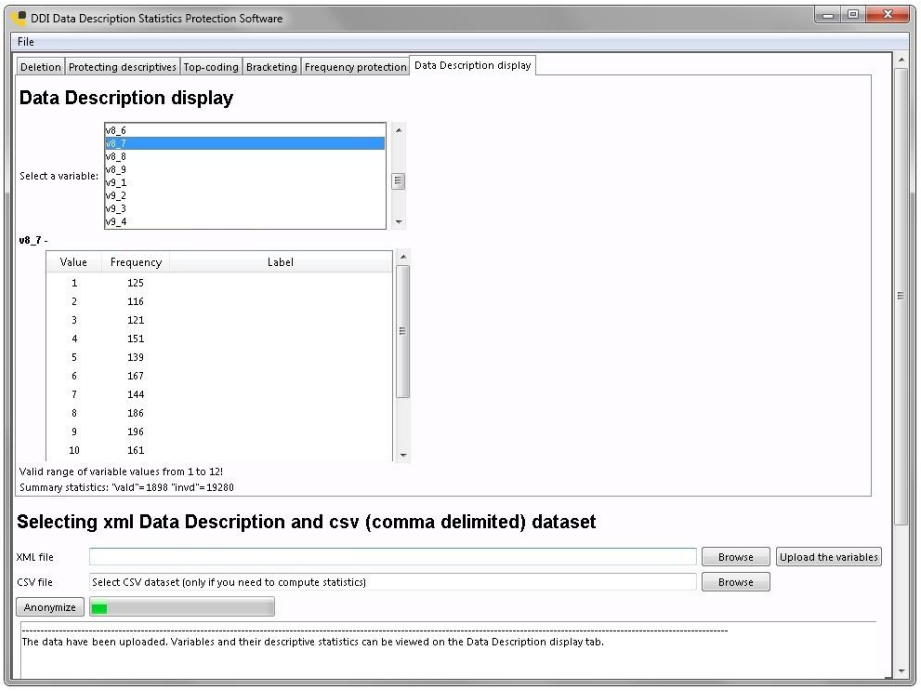

Fig. 6. Data Description display tab

For archiving purposes and due to the fact that there might be several people working on the Data Description protection, with statistical office employee(s) always being the one(s) to check the final version before the distribution, the tool was programmed to create a full anonymization process report with all the techniques used listed and changes made to the original Data Description XML file explained in detail. The report is added next to the anonymized XML version in the TXT format. The practice of creating data protection report is similar to the established practice of other anonymization tools, such as μ-ARGUS [6], but is in this case adjusted to the specifics of the DDI Data Description protection.

Last but not least, although that using the presented software should be a very simple task for users with previous anonymization experience, PDF guidelines were nevertheless written and added to the File drop-down menu. The guidelines include:

- detailed instructions on how to use individual anonymization techniques
- instructions on how to prepare uploaded CSV and XML files
- a list of common errors with instructions on how to fix them

5 Concluding remarks

In this paper we have presented the reasons why the DDI Data Description Statistics Protection Software needed to be developed, as well as the mixture of anonymization methods and techniques which were adjusted to the specifics of the XML Data Description code and descriptive statistics display. Some functions were integrated into the tool without the immediate or desperate need for them, but with potential future issues regarding statistical disclosure control in mind; despite the fact that many people believe that the aggregated data anonymization of this kind is unnecessary and how aggregating data is an anonymization technic itself. However, they have to be aware of how strictly national statistics offices have to work in accordance to the law and how important trust in official statistics is. Therefore each time any kind of data, either microdata or tabular and aggregated data are made publically available to either provide information to the public or for promotional purposes, a lot of caution is required and attention to detail paid. Therefore we believe that the tool presented in this paper serves and will serve the purpose, most certainly in the Slovenian official statistics environment.

There is also one notable avenue for future work. As the tool is designed for use with the Nesstar Publisher exported DDI Data Description XML files, it could be relatively easily re-programmed to work with that software and not as an independent follow-up program. One of the options we have in mind would be to programme it as an add-on which would perform all of the anonymization operations, while Nesstar Publisher would export a clean, adequately protected Data Description XML file.

References

1. Cornell Anonymization Toolkit. http://anony-toolkit.sourceforge.net/ (2009). Accessed 4 July 2014
2. Data Documentation Initiative: XML Schema Outline -- Version 2.1. http://www.ddialliance.org/Specification/DDI-Codebook/2.1/DTD/Documentation/DDI2-1-tree.html (2014). Accessed 4 July 2014
3. Data without Boundaries. http://www.dwbproject.org/ (2014). Accessed 4 July 2014
4. Hundepool, Anco et al.: Statistical Disclosure Control. John Wiley & Sons Inc., Chichester (2010)
5. Hundepool, Anco et al.: τ-ARGUS User's Manual Version 3.5. Statistics Nederlands, The Hague (2011)
6. Hundepool, Anco et al.: μ-ARGUS User's Manual Version 4.2. Statistics Nederlands, The Hague (2008)
7. International Household Survey Network: Reducing the disclosure risk. http://www.ihsn.org/home/node/201 (2014). Accessed 4 July 2014
8. Inter-university Consortium for Political and Social Research (ICPSR). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). ICPSR, Ann Arbor (2012)
9. NSD: Nesstar Publisher. http://www.nesstar.com/software/publisher.html (2014). Accessed 4 July 2014
10. NumPy. http://www.numpy.org/ (2014). Accessed 4 July 2014

11. Python. https://www.python.org/ (2014). Accessed 4 July 2014
12. SciPy. http://www.scipy.org/ (2014). Accessed 4 July 2014
13. Social Science Data Archives. About the data of authorized producers of national statistics. http://www.adp.fdv.uni-lj.si/eng/projekti/uradne-statistike/ (2014). Accessed 4 July 2014
14. Templ, M.: Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. Transactions on Data Privacy 1(2), 67-85 (2008)
15. United Nations Economic Commission for Europe. Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice. United Nations, New York and Geneva (2007)