Project number:	317871
Project acronym:	BIOBANKCLOUD

Project title: Scalable, Secure Storage and Analysis of Biobank Data

Project website URL: http://www.biobankcloud.com/

Project Coordinator Name and Organisation:
Jim Dowling, KTH

E-mail: jdowling@kth.se

WORK PACKAGE 5 :

Data-intensive computing with genomic data

Work Package Leader Name and Organisation:
MICHAEL HUMMEL, Charité-Universitätsmedizin Berlin (CHARITÈ)

E-mail: michael.hummel@charite.de

PROJECT DELIVERABLE

D5.2: *K*-anonymization tools

Deliverable Due date (and month since project start): 2013-11-30, m12

Deliverable Version: v0.2

Document history

Version	Date	Changes	By	Reviewed
0.1	2013-11-26	First draft	Lora Dimitrova Ali Gholami	Michael Hummel
0.2	2013-11-29	Final version	Lora Dimitrova Ali Gholami Roxana Merino Martinez Jim Dowling	Michael Hummel

Abstract

For privacy protection of the individuals-subjects of released microdata, data are often de-identified, due to laws and legislations such as European personal data protection directive, HIPPA and so on. De-identification means that identifiers like names and social security numbers are removed or encrypted, but de-identification is not sufficient to anonymize these data. Microdata often contain other data, which can be linked to publicly available data and thus the individuals behind the data can be re-identified.

In deliverable D5.2 "*K*-anonymization tools" we will describe *k*-anonymity as a concept for protection of microdata and also different tools for *k*-anonymization in detail that is required to be adopted in the BiobankCloud.

Table of contents

1. Introduction

Agencies and many other organizations publish microdata about individuals for purposes like demographic and public health research. Although attributes like name and social security number are removed from the database (de-identification), there is still the possibility to identify de-identified individuals from the database. By linking the de-identified database to other publicly available databases and using attributes like e.g. ZIP code, sex and date of birth, individuals can be re-identified (1).

In deliverable D5.2 we describe how the privacy of individuals-subjects of released microdata can be threatened and suggest different techniques and tools that are suitable for privacy protection.

In section 2 we show an example how microdata can be attacked and individuals re-identified, and describe two different approaches for privacy protection, namely *k*-anonymity (section 2.1) and *l*-diversity (section 2.2). In section 3 we give a detailed description of four distinct *k*-anonymization tools, which can be used for microdata protection: UTD Anonymization ToolBox (section 3.1), TIAMAT (section 3.2), ANON tool (section 3.3) and the eCPC toolkit (section 3.4). Section 4 contains the conclusion of our analyses.

2. Privacy protection of published microdata

Microdata are tables containing unaggregated information like medical, census, voter registration or customer data, about individuals. The entities, to which these microdata refer are called respondents and their anonymity must be protected (2, 3). Anonymity is defined in two ways (4). The first definition is:

“Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set”

The second definition is from the attacker’s perspective:

“Anonymity of a subject from an attacker’s perspective means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set”

For protecting the privacy of the respondents the microdata are de-identified, meaning that identifiers such as names, addresses, phone and social security numbers are often removed or encrypted. However de-identification does not guarantee the privacy of the respondents. The published microdata include quasi-identifiers (QIDs) such as set of attributes including gender, date of birth, ZIP code, race, which linked to publicly available data (local census data, voter lists, city directories, data from motor vehicle agencies, tax assessors, real estate agencies etc.) can lead to the unique identification of the respondents in the population (2, 3). How such a linking attack may compromise the anonymity of an individual is demonstrated exemplarily in Ciriani et al. (2007) (2). In the upper table of Figure 1 medical data, which are de-identified by removing names and social security numbers, is shown. Other data like race, date of birth, sex, ZIP code and marital status are maintained. These QIDs can be linked to voter lists and an example for a voter list is shown in the second table of Figure 1. The voter list contains additional information like name, address and city. In the medical data there is only one person, who is born on 64/04/12, female, divorced and living in the 94142 area, and if there is a unique match in the voter list, this person can be identified as Sue J. Doe, living in 900 Market Street / San Francisco, suffering from hypertension.

SSN	Name	Race	Date of birth	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Fig. 1. De-identified private table (medical data)

Name	Address	City	ZIP	DOB	Sex	Status
.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....
Sue J. Doe	900 Market St.	San Francisco	*94142*	*64/04/12*	*F*	*divorced*
.....	.....	.....	.....	.....	.....	.....

Fig. 2. Non de-identified public available table

Figure 1: Example for released de-identified medical data and publicly available data (voter list). Tables from Ciriani at el. (2).

Different approaches have been developed for the protection of de-identified microdata from linking attacks. In summary, anonymization methods can be classified as described in Fung et al. (2010) (5):

- Generalization and suppression: to hide details of QIDs with each operation.
- Anatomization and permutation: to de-associate the QID from the sensitive attributes.
- Random perturbation (additive noise, data swapping, synthetic data generation): to replace the original value with a perturbed value. But this method degenerates the quality of the published data because of the additive noise and other statistical parameters.

2.1 *k*-Anonymity

To provide protection against attribute linkage, Latanya Sweeney (6) proposed *k*-anonymity as a model for privacy preserving of QIDs. The *k* value is the minimum

number of the records in a table that have similar QIDs. This notion of *k* records in a group reduces the risk of re-identification of a participant to the probability of 1/*k*. However, *k*-anonymity is weak regarding the background knowledge of the adversary about a victim as described in Machanavajjhala et al. (2007) (3).

k-Anonymization works in the following way: first all fields containing QIDs in the microdata are identified and their content is generalized, suppressed or both. While suppression means that individual attributes are replace with a *, generalization means that individual attributes are replaced with a broader category (e.g. replacing the age by an age group). Thus in the final *k*-anonymized microdata for each record there are at least *k-1* other records, from which it can't be distinguished (7).

2.2 *l*-Diversity

To overcome the limitations of *k*-anonymity, *l*-diversity can provide an extra privacy measure to protect the anonymity of the individuals from re-identification through the adversary's background knowledge (3). The value of *l* defines at least *l* "well-represented" sensitive values in the table to reduce the confidence of inferring a sensitive attribute within a group. Entropy (E) of the entire table must hold $E \geq \log(l)$ to ensure every distinct QID block, at least has *l* distinct values for the sensitive attribute.

3. Anonymization tools

Different tools have been developed for privacy protection of microdata and such anonymization tools are described in the next section.

3.1 Anonymization ToolBox (UT Dallas)

The UTD Anonymization ToolBox (http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home) was developed by the Data Security and Privacy Lab (The University of Texas at Dallas, Dallas, TX, USA) and the toolbox as well as the documentation can be downloaded from the Lab's website.

The input files supported are unstructured text files. The output files have the same format by default (*genVals*) and they are differing from the input files only in

generalization of specific values of the QID attributes. Two additional output formats are supported: *genValsDist*, which contains additional equivalence-wide statistics, and *anatomy*. When the *anatomy* format is chosen, two datasets are release after anonymization. The first is consisting of the QIDs replaced by an equivalence class, i.e. a set of records with the same QID values. The second dataset is containing the same new attribute and additionally the ground level QID values.

The configuration file is in XML format and the root node *config* has no more than 5 children:

- *input*, containing the input parameters.
- *output*, defining the output format. This child is optional.
- *id*, which is a list of identifier attributes and is optional too.
- *qid* is a list of QID attributes.
- *sens* is an optional child and contains a list of sensitive attributes.

With help of the configuration file almost every anonymization parameter can be specified. In Table 1 the parameters that can be set through program arguments are shown. The first column contains the parameter keys, the second the values, which the toolbox will recognize, and the third the default values of the parameters.

Parameter	Domain or explanation	Default
-method	Choose from {Datafly, Incognito_k, Incognito_l, Incognito_t, Mondrian, Anatomy}	Mondrian
-config	path to the configuration file	config.xml
-k	any positive integer value	10
-l	any positive real value	10
-c	any positive real value	0.2
-t	any value within the range (0, 1)	0.2
-suppThreshold	max. number of tuples to be suppressed (applies only to Datafly and Incognito_K)	k
-input	input filename	N/A
-separator	character sequence separating attribute values	comma
-output	output filename	N/A
-outputFormat	Choose from {genVals, genValsDist, anatomy}	genVals

Table 1: Command line parameters. Table from http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=doc.

Six distinct anonymization methods can be performed with the toolbox:

- **Datafly**: a full-domain generalization algorithm. Generalization is performed until every combination of QID values is represented at least *k* times. Datafly can also perform suppression. The disadvantage of this method is that it tends to over-generalize the data.
- **Mondrian Multidimensional *k*-Anonymity**: an anonymization method that performs generalization at the equivalence level thus solving the problem with over-generalizing data. The result is an anonymized dataset that consists of more equivalence classes and the released information on QID attributes is more accurate.
- **Incognito**: a minimal full-domain generalization algorithm developed by LeFevre et al. (2005) (1). The method can perform generalization as well as suppression.
- **Incognito with *l*-diversity**: *k*-anonymity can be attacked by adversaries possessing background knowledge on individuals included in the dataset. *l*-Diversity diversifies sensitive values within each equivalence class thus defending the microdata against such attacks.

- **Incognito with t-closeness**: a privacy notion proposed by Li et al. (2007) (8), which tries to overcome the limitations of *l*-diversity. The requirements of the approach are that the distribution of a sensitive attribute in every equivalence class is close to its distribution in the overall table.
- **Anatomy**: an algorithm that generates *l*-diverse equivalences, developed by Xiao and Tao (2006) (9). Anatomy is overcoming the disadvantages of generalization by preserving privacy **and** correlation in the microdata.

The UTD ToolBox can be extended with further anonymization methods and the developers have planned to update it frequently.

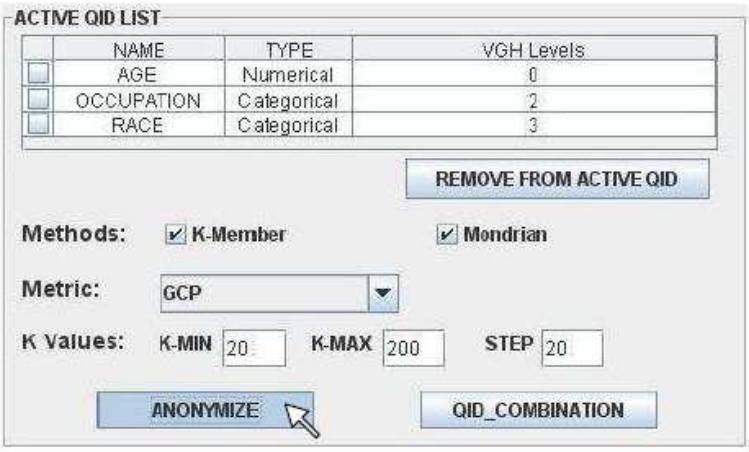
3.2 TIAMAT

With TIAMAT (10) different *k*-anonymization techniques can be compared with respect of their accuracy and runtime performance, and appropriate settings of the parameters for anonymization can be found. The main features of the tool are demonstrated in the next section by taking the example of anonymizing the Adult database from the UC Irvine repository, containing 30162 records.

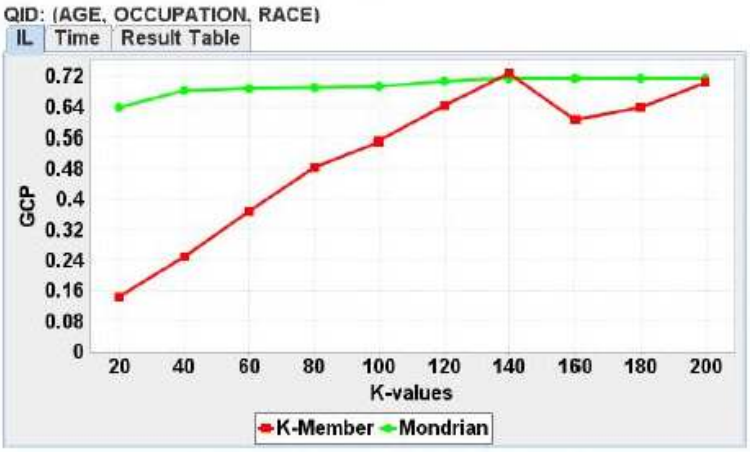
TIAMAT automatically populates candidate QID lists from SQL-compliant databases. These lists include not only the type and the domain of the QIDs, but also the number of levels in the taxonomy tree (i.e. the Value Generalization Hierarchy (VGH)), which is associated to the respective QID. The VGH can be visualized and edited with help of the VGH Editor tool (see Figure 2).

Figure 2: Candidate QID list generated by TIAMAT. Table from Dai et al. (2009) (10).

Two kinds of analysis are supported by TIAMAT: the Comparative Analysis of Anonymization Techniques and the QID Change-Impact Analysis (for choosing a set of QID attributes). The first analysis is used for head-to-head performance comparison of anonymization techniques. Here the user must choose a fixed QID attribute combination and the tool compares the data accuracy and execution time of distinct anonymization techniques. In Figure 3 an example is shown. The attributes chosen as QIDs are Age, Occupation and Race and the Mondrian and *k*-Member anonymization methods can be compared. The tool allows the user also to select the *k*-range and the intervals between the *k*-values. Additionally two different metrics are available for evaluating information loss (i.e. data accuracy): Global Certainty Penalty (GCP) and Classification Metric (CM) (see Figure 3a).



(a) Choosing Anonymization Parameters



(b) Information Loss Visualization Panel

AGE	WORKCLASS	FNLWGT	EDUCATION	MARITAL_STATUS	OCCUPATION	RELATIONSHIP
25.0~36.0	Private	211266.0	Some-college	other	Occupation	Other-relative
25.0~36.0	Private	189775.0	Some-college	other	Occupation	Own-child
25.0~36.0	Private	212563.0	Some-college	other	Occupation	Unmarried
25.0~36.0	Private	70282.0	Some-college	other	Occupation	Unmarried
29.0~47.0	Private	88419.0	education	Never-married	Exec-managerial	Not-in-family
29.0~47.0	Private	163003.0	education	Never-married	Exec-managerial	Other-relative
29.0~47.0	Private	348491.0	education	Never-married	Exec-managerial	Not-in-family
29.0~47.0	Private	209734.0	education	Never-married	Exec-managerial	Unmarried
29.0~43.0	Private	419721.0	HS-grad	Never-married	Other-service	Unmarried
29.0~43.0	Private	206365.0	HS-grad	Never-married	Other-service	Not-in-family
29.0~43.0	Private	39581.0	HS-grad	Never-married	Other-service	Not-in-family
29.0~43.0	Private	70240.0	HS-grad	Never-married	Other-service	Own-child

(c) Anonymized Table (*k*-Member)

Figure 3: Comparative analysis of anonymization techniques with TIAMAT. Figures from Dai et al. (2009) (10).

The results of the anonymization are shown in Figure 3b and they are presented in three different ways. A summary plot of information loss (IL tab) and execution time (Time tab) is available along with the Result Table (see Figure 3c), in which the

contents of the anonymized tables for each *k*-value are shown. The anonymized tables can be stored back to the database.

The comparison of both anonymization techniques for the test data reveals that for most of the *k*-values the *k*-Member method is superior to the Mondrian as to information loss. But for some *k*-values like *k* = 140, there is an increase of the GCP and for determination of the QID attributes, that cause this increase, the QID Change-Impact Analysis tool can be used.

This feature allows the user to evaluate different QID attribute sets and to choose the one, which is leading to the most minimal information loss. For this purpose the QID_Combination feature (see Figure 3a) must be used and as a result the tool combines the QIDs from the active QID list in all possible ways (see Figure 4) and anonymizes the data respectively.

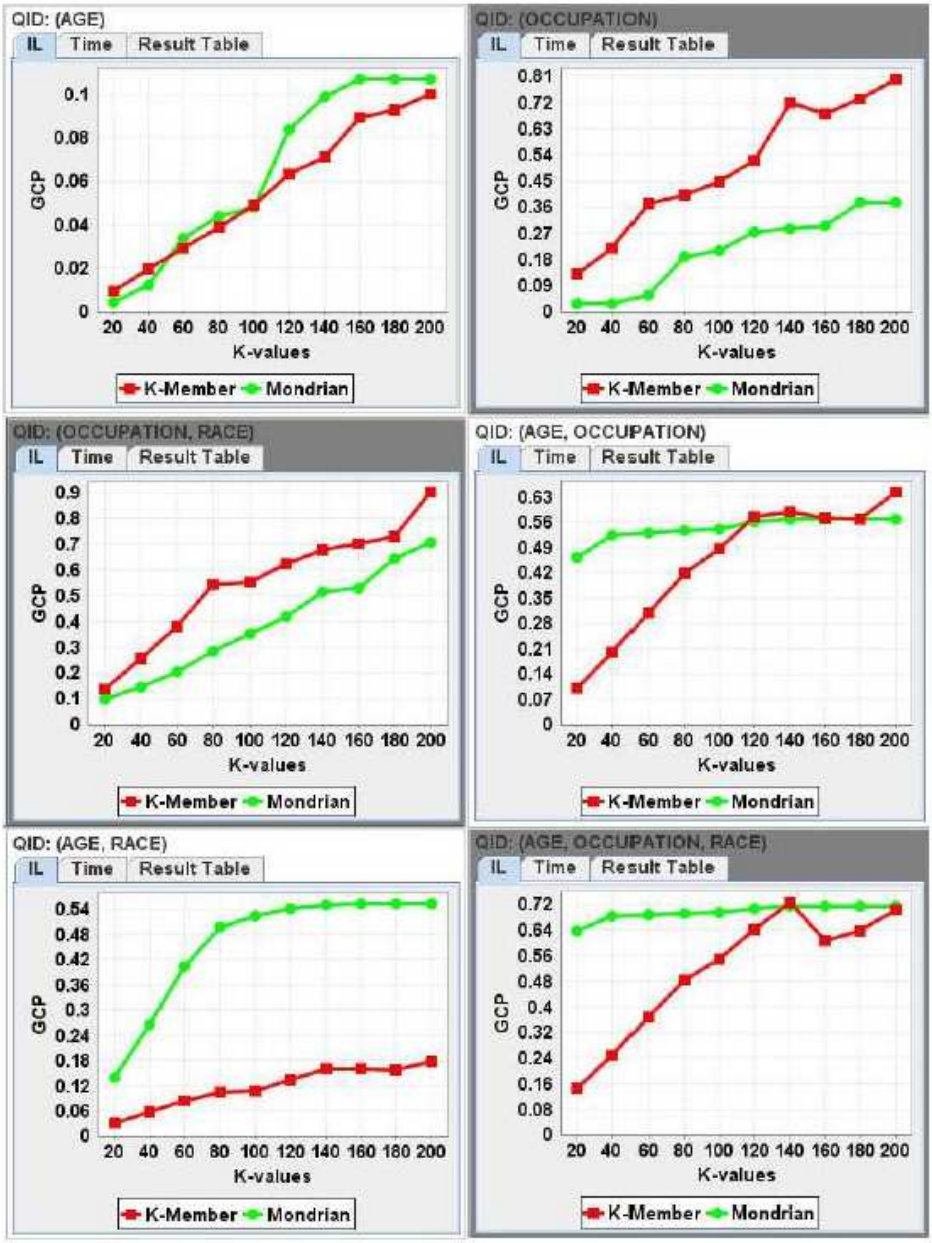


Figure 4: QID Change-Impact Analysis with TIAMAT. Figures from Dai et al. (2009) (10).

Depending on the result, the most suitable QID set can be selected for the final anonymization of the data.

3.3 ANON tool

The ANON tool (11) was developed by Dr. Johann Eder (Alpen Adria University, Klagenfurt, Austria) for the TMF (TMF–Technology, Methods, and Infrastructure for Networked Medical Research; Berlin, Germany), which is Germany's umbrella organization for networked medical research. The ANON tool is adding *k*-anonymity and *l*-diversity to a given data set while keeping data loss minimal.

The software takes different input files: XML-files, which contain a description of the configuration parameters or data from a data source like a JDBC connection (i.e. data from a table/view/query from any database), a XML or a CSV file (i.e. plain text consisting of records, which are divided into fields separated by commas). It generates two output files: a report XML-file or a CSV / XML file or a table in a database, which contains the computed *k*-anonymous and *l*-diverse data set. A schematic overview of the input and output files of the ANON tool is shown in Figure 5.

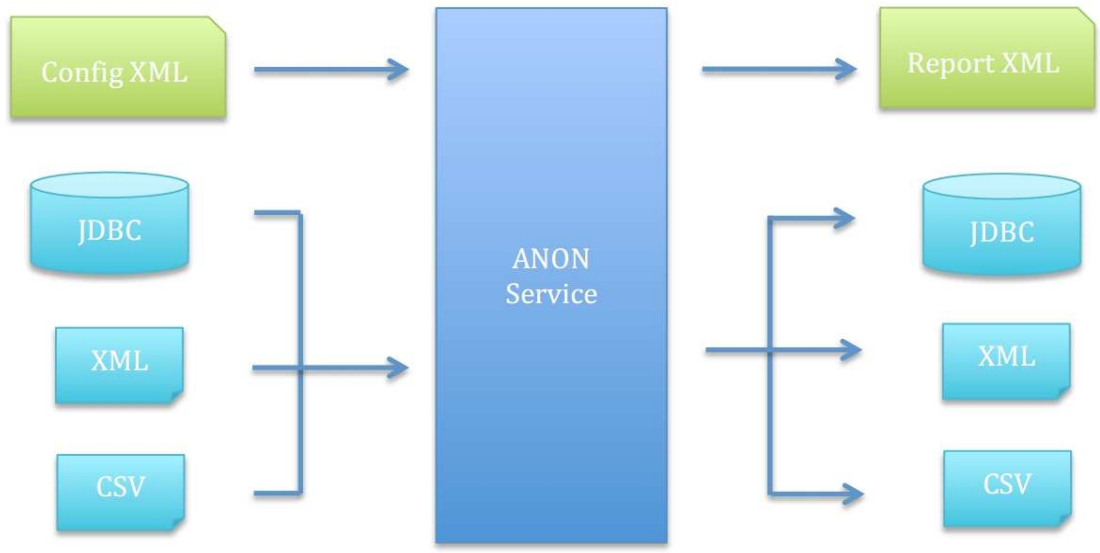


Figure 5: Scheme of the inputs and outputs of the ANON tool. Scheme from the ANON manual (11).

There are three different distributions of the ANON tool. ANON_UI.jar has a simple graphical user interface (see Figure 6) and is a java application, which should be run with Java version 1.6 or greater.

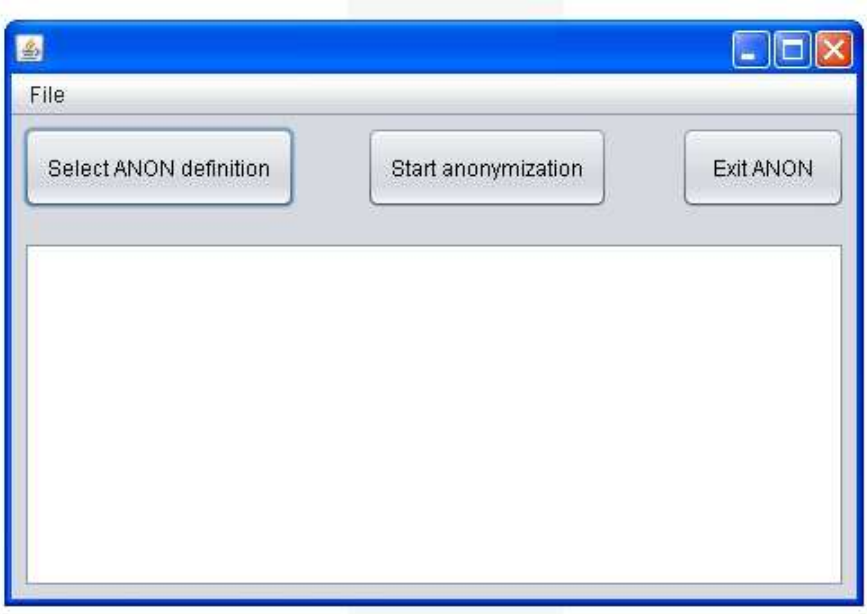


Figure 6: Graphical user interface of ANON_UI.

ANON.war is a web service application, which is running on a GlassFish Application Server. This application requires the GlassFish Application Server (version 3.1 or greater) and also a Java Development Kit (JDK7 or greater).

Additionally all source files for the UI application and the web service can be downloaded. The ANON Definition file (ANONSchema.xsd) contains five sections, which will be described in detail in the following section:

- **GeneralizationHierarchies**: This section may consist of sequences of “NumericalHierarchy” and “CategoricalHierarchy” definitions and it defines the value generalization hierarchies, which are used for the anonymization. The first one is a numerical hierarchy, in which the intervals of a certain level can differ, while the second one defines hierarchies for categorical attributes.

- **Parameters**: This section defines the anonymization settings and has four sections:
  - **kValue**, which defines the parameter *k* for *k*-anonymization.
  - **Threshold**, which specifies the fraction of records for suppression, meaning that these records may be removed from the result.
  - **SearchType**, which defines the search algorithm for anonymization.
  - **WorkReport**, which defines the parameters for the anonymization report. In this section the location for saving of the report is being specified. The report provides information about errors or problems in the ANON Definition file or statistical information about the anonymization of the input data with the ANON tool.

- **AttributesDefinition**: This section defines the attributes from the input data, which should be read, and the way of their handling. The AttributesDefinition consists of the following sections:
  - **annonymizationType**, which defines the anonymization type of the attribute. In this section the values, which can be chosen, are: **l-attribute** for sensitive attributes that must be *l*-diverse; **k-attribute** for QIDs that must be *k*-anonymous; **dontcare** for attributes that are published as they are; **ignore** for attributes that disappear from the result.
  - **useGeneralizationHierarchyWithID**, which is specifying a generalization hierarchy and is needed in case of the “k-attribute” annonymizationType.
  - **Label**, which defines the attribute’s name.
  - **SQLName** defines the name of the attribute, which doesn’t differ from its name in the database.
  - **Limit**, which defines the maximal generalization level of a value and concerns only attributes with “k-attribute” annonymizationTypes.
  - **Priority** defines the generalization of attributes with low priorities before generalization of attributes with high priorities.

- **DatasourceDefinition**: This section contains one or several “Source” sections from the CSVSource, XMLSource or JDBCSource type and defines the source of the data for anonymization.
- **OutputDefinition**: This section consists of only one “OutputTo” section, which may be a JDBC, CSVFile or a XMLFile, and defines the target for saving of the anonymized data.

The ANON tool is an open source tool and it can be downloaded together with its documentation from the TMF website (http://www.tmf-ev.de/Produkte/Uebersicht. aspx# P100201).

3.4 eCPC toolkit

In the BiobankCloud we will use an anonymization toolkit, which is developed in the eScience for Cancer Prevention and Cure (eCPC; http://www.e-science.se/ community/eCPC) project. The eCPC is a flagship project within the Swedish e-Science Research Center (SeRC), aiming to develop a modular system for prediction of cancer initiation and progression using modeling and simulation. An important part of the project is to integrate data from different sources, such as biobanks containing data about samples, and clinical health registries (quality registries) containing information about patients and their diseases, treatments, and outcomes. This integrated data can then be used in subsequent modeling and simulation efforts. A big hurdle in medical data integration is the acceptance and participation of data providers. The eCPC provides a toolkit which lowers the barriers for data providers to participate through a portable user interface, which is implemented in Java.

The eCPC toolkit produces tabular microdata in comma separated values (CSV) format from relational databases using java database connectivity (JDBS), using data management component, as shown in Figure 7.

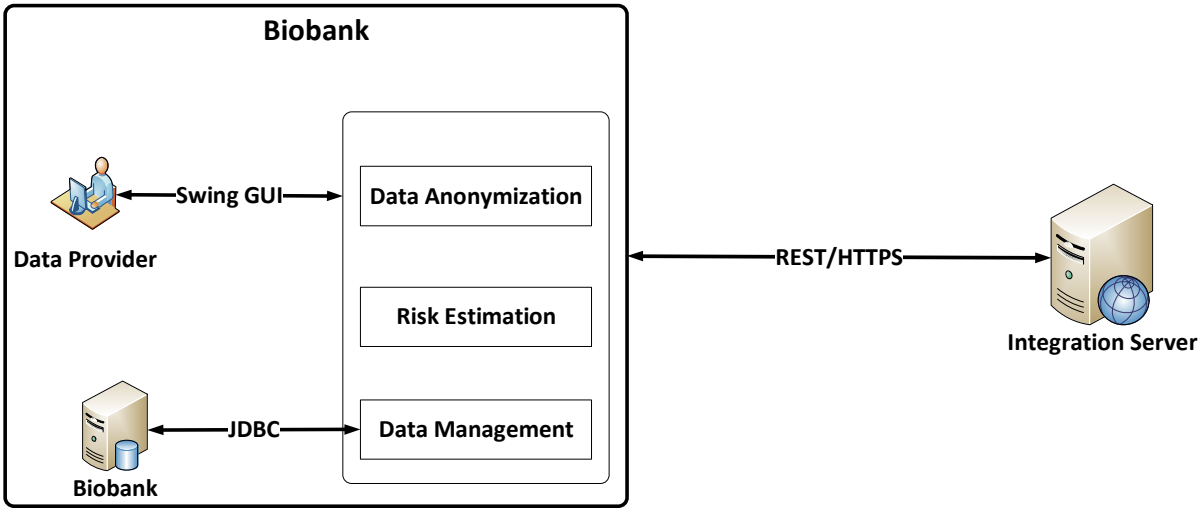


Figure 7: Architecture of the eCPC toolkit.

A user can login to the toolkit and after successful authentication, will be able to anonymize the CSV files according to *k*-anonymity and *l*-diversity algorithms.

Furthermore, the toolkit provides functionality of risk assessment for a selected sample microdata in CSV format. For instance, the data provider selects a set of key attributes to assess the risk of data publishing and only if the risk is lower than a threshold, he/she decides to publish the data.

The eCPC toolkit graphical user interface (GUI) is implemented in Java and the anonymization algorithms are implemented using sdcMicro (http://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf) which is a R package to anonymize microdata. However, due to compatibility issues of R with Java, the toolkit runs the anonymization algorithms in batch mode within the R environment.

As an extra security measure, the eCPC toolkit provides encryption of the anonymized data sets for cross-query purposes across different data providers. As first step, the direct personal identifiers will be hashed using secure hash function SHA-512 and the hashed personal identifiers then will be encrypted using advanced encryption standard (AES), as shown in Figure 8, where the data providers have

access to the encryption keys. The two level encryption makes it very difficult for an adversary to read and make inference attacks on the anonymized data sets.

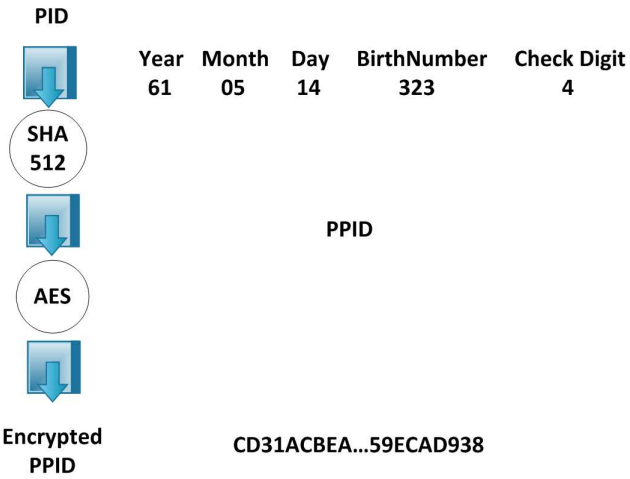


Figure 8: De-identification and encryption of anonymized attributes in the eCPC toolkit.

The eCPC pseudonymity model extracts and converts the personal numbers using a two-level mechanism that maintains the possibility of joint queries over anonymous records in different collections. In Sweden, a personal number (Swedish civic registration number) is a 10-digit PIN issued by the National Tax Board for all residents in the country. The personal number or personal identifier (PID) is structured in three parts: date of birth, a three-digit birth number and a check digit. The date of birth construction contains two-digits each for the year, month, and date of birth, e.g., 610514. This is followed by a three-digit birth number (e.g., 323) and a check digit (e.g., 4), as shown in Figure 8. The birth number value will be a number between 001 and 999, where the last digit is also used to indicate the gender, with men given an odd number and women an even number.

In summary, to ensure the privacy of sensitive patient data, the eCPC toolkit applies the guidelines for safe microdata, outlined as follows:

- The eCPC toolkit removes all explicit identifiers, and it will extract and de-identify all the PIDs, as shown in Figure 8;

- Then it categorizes the remaining attributes to determine the key variables according to both legal requirements and domain specific judgments, which may be subjective. Key variables might also be split into further categories according to the level that they are identifying. This distinction is useful when prioritizing which variables need to be modified to enhance safety: more identifying keys are modified first for observations that have a considerable high risk. This graded approach allows for better data quality preservation and therefore higher data utility. Figure 9, shows the anonymization window, for a set of key attributes (QIDs) for *k*-anonymization and also sensitive attributes to ensure *l*-diversity;

Figure 9: *k*-Anonymization and *l*-diversity check in the eCPC toolkit.

- When key variables have been identified, the re-identification risk needs to be assessed. This is done by looking at the uniqueness of the observed entries through frequency counting and calculating probability estimates based on extrapolating models taking population frequencies into account, as shown in Figure 10;

Figure 10: Re-identification risk calculation of microdata to be published.

- Entries that stand out from the rest and therefore have a considerable risk to be subject to re-identification are then modified. Numerous modification algorithms exist, namely generalization or global recoding and local suppression of outstanding values, recoding, swapping, rank swapping or perturbing with post randomization – not to be confused with randomized questionnaires when collecting data, hence the name post randomization.
- The methods to be applied depend on the nature of the variables, whether they are categorical or continuous, their structure such as significance order, hierarchy, geography, semantics and the size of the dataset in question. Nevertheless, each algorithm applied is recorded in a logbook for the analyst's documentation, e.g., the nature of the added noise, if any;
- The re-identification risk has to be measured again and the information loss has to be evaluated. If the risk is deemed acceptable and the quality of the data remains adequate, then the resulting microdata can be considered as safe. If not, the previous step needs to be repeated;
- Finally, the data provider e.g., Biobank, publishes the pseudonymized data sets to the integration server.

As we discussed above, the eCPC toolkit seems to provide a usable and portable platform, however it may need to be adopted according to the BiobankCloud requirements. For instance, the Java swing in the toolkit should be replaced with a web interface to be integrated with the BiobankCloud portal. The toolkits backend also may require few modifications to be working with the BiobankCloud platform.

4. BiobankCloud anonymization architecture

For anonymization of the microdata, we propose an anonymization architecture, as sketched in Figure 11. The anonymization toolkit is composed of three main components: *k*-anonymizer, risk calculator and data publisher. The *k*-anonymizer component, loads a CSV file and presents the metadata for the user. A set of attributes will be used for anonymization with the default values of $k = 3$ and $l = 3$, respectively for *k*-anonymity and *l*-diversity. The risk calculator provides re-identification risk of individuals for data publishing. A user selects the anonymized CSV file and measures the re-identification risk based on some attributes. Finally, the data publisher will export the anonymized data using secure channels and REST API.

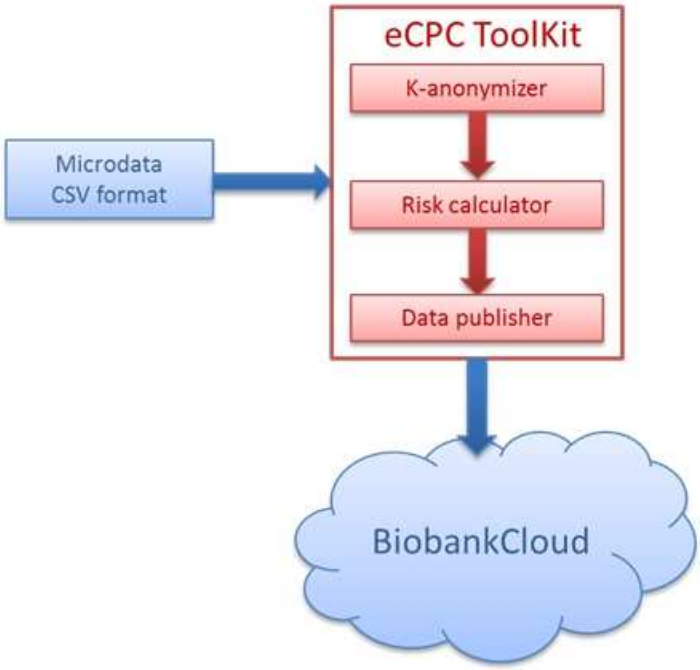


Figure 11: Anonymization architecture of the BiobankCloud microdata.

5. Conclusions

Microdata are tables containing unaggregated information about individuals and the anonymity of these individuals must be protected. *k*-Anonymity is proposed as a model for privacy preserving and in deliverable D5.2 we have described different tools for *k*-anonymization of microdata. In the BiobankCloud platform we will integrate and use the eCPC toolkit developed in the eScience for Cancer Prevention and Cure project for anonymization of microdata.

References

1. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain K-anonymity. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* 2005**:** 49-60.

2. Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P. k-Anonymity. *Advances in Information Security* 2007.

3. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-Diversity: Privacy Beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data* 2007; **1**.

4. Pfitzmann A, Hansen M. A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. 2010.

5. Fung BCM, Wang K, Fu AW-C, Yu PS. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques. 2010.

6. Sweeney L. Simple Demographics Often Identify People Uniquely. 2000.

7. Williams R, Blum M. K-Anonymity. *Summer REU (Research Experiences for Undergraduates)* 2007.

8. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference* 2007**:** 106-115.

9. Xiao X, Tao Y. Anatomy: simple and effective privacy preservation. *VLDB '06: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment* 2006**:** 139-150.

10. Dai C, Ghinita G, Bertino E, Byun JW, Li N. TIAMAT: a Tool for Interactive Analysis of Microdata Anonymization Techniques. *Proceedings of the VLDB Endowment* 2009; **2**(2).

11. Eder J, Koncilia C, Ciglic M. ANON Manual. 2013.