Project no. 826278

SERUMS

Research & Innovation Action (RIA)
SECURING MEDICAL DATA IN SMART-PATIENT HEALTHCARE SYSTEMS

Report on Initial Data Masking, Data Fabrication and Semantic-Preserving Encryption
D4.1

Due date of deliverable: 31st December 2019

Start date of project: 1st January 2019

Type: Deliverable
WP number: WP4

Responsible Institution: IBM
Editor and editor's address: Michael Vinov (vinov@il.ibm.com)
Partners Contributing: UCL, USTAN, SCCH, ZMC, UCY, FCRB

Approved by:

Version 1.0

Project co-founded by the European Commission within the Horizon H2020 Programme		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Release History

Release No.	Date	Author(s)	Release Description/Changes made
V0.1	01/12/19	Michael Vinov (IBM)	Initial draft
V0.2	12/12/19	Thomas Given-Wilson (UCL)	
V0.3	12/12/19	Michael Vinov (IBM)	Added Sections on Semantic-preserving Data Encryption and synthetic data fabrication
V0.4	17/12/19	Michael Vinov (IBM)	Updated the Introduction and Conclusions sections
V0.5	30/12/19	Juliana Bowles (USTAN)	Updated partners, minor corrections

SERUMS Consortium

Partner 1	**University of St Andrews**
Contact Person	Name: Juliana Bowles Email: jkfb@st-andrews.ac.uk
Partner 2	**Zuyderland Medisch Centrum**
Contact Person	Name: Mark Mestrum Email: m.mestrum@zuyderland.nl
Partner 3	**Accenture B.V.**
Contact Person	Name: Bram Elshof, Wanting Huang Email: bram.elshof@accenture.com, wanting.huang@accenture.com
Partner 4	**IBM Israel Science & Technology Ltd.**
Contact Person	Name: Michael Vinov Email: vinov@il.ibm.com
Partner 5	**Sopra-Steria**
Contact Person	Name: Andre Vermeulen Email: andreas.vermeulen@soprasteria.com
Partner 6	**Université Catholique de Louvain**
Contact Person	Name: Axel Legay Email: axel.legay@uclouvian.be
Partner 7	**Software Competence Centre Hagenberg**
Contact Person	Name: Michael Rossbory Email: michael.rossbory@scch.at
Partner 8	**University of Cyprus**
Contact Person	Andreas Pitsillides Email: andreas.pitsillides@ucy.ac.cy
Partner 9	**Fundació Clínic per a la Recerca Biomèdica**
Contact Person	Name: Santiago Iriso Email: siriso@clinic.cat

Table of Contents

RELEASE HISTORY 2

SERUMS CONSORTIUM 3

EXECUTIVE SUMMARY 5

1 INTRODUCTION 6

1.1 ROLE OF THE DELIVERABLE 6
1.2 RELATIONSHIP TO OTHER SERUMS DELIVERABLES 6
1.3 STRUCTURE OF THIS DOCUMENT 6

2 DATA MASKING AND SYNTHETIC DATA FABRICATION 7

2.1 INTRODUCTION TO DFP 7
2.2 SYNTHETIC DATA FABRICATION USING CSP 8
2.3 FABRICATION OF SYNTHETIC HEALTHCARE DATA 10
2.3.1 ZUYDERLAND MEDISCH CENTRUM SMART HEALTH CENTRE (ZMC) USE CASE 10
2.3.2 HOSPITAL CLINIC OF BARCELONA SMART PLATFORM (FCRB) USE CASE 11
2.3.3 PERSONAL CANCER TREATMENT (USTAN) USE CASE 13
2.4 DATA MASKING 13

3 VERIFICATION OF FABRICATED DATA QUALITY 15

4 SEMANTIC-PRESERVING DATA ENCRYPTION 16

5 CONCLUSIONS 18

REFERENCES 19

Executive Summary

Securing Medical Data in Smart Patient-Centric Healthcare Systems (SERUMS) is a research project supported by the European Commission (EC) under the Horizon 2020 program. This document is the first deliverable of Work Package 4: "Secure and Privacy-Preserving Data Communication". The leader of this work package is IBM, with involvement from the following partners: UCL, USTAN, SCCH, ZMC, UCY and FCRB. The goal of this work package is to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart healthcare system. The objectives of WP4 are to:

- develop advanced data masking and synthetic data fabrication technologies to enable sharing of personal medical data between components of the Smart Health Centre system developed in WP6;
- develop metrics and techniques to verify both the security and the functional properties of the advanced data analytics and the Serums patient-centric Smart Health Centre system;
- explore and develop technology for encrypting information while preserving certain required semantics, in order to enable advanced data analytics while adhering to privacy regulations.

This deliverable entitled "Report on Initial Data Masking, Data Fabrication and Semantic-Preserving Encryption" is the first deliverable of the WP4. It describes initial versions of the data masking and data fabrication technologies that are used in the project to enable sharing of personal healthcare data between the project partners and development of the Smart Health Centre. The deliverable report also describes an initial version of the technology to verify the quality of fabricated synthetic data on initial versions of the data analytics and authentication tools and an initial version of the semantic-preserving data encryption technology to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

1 Introduction

1.1 Role of the Deliverable

The aim of this deliverable is to report and describe the design and development of initial versions of the data masking, data fabrication, data quality verification and semantic-preserving data encryption technologies. All these technologies are used to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart healthcare system and to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

1.2 Relationship to Other SERUMS Deliverables

Tasks 4.1 and 4.2 of WP4 are closely related to the work done in WP2 – "Smart Patient Record Construction". Masked data and synthetic fabricated data of WP4 is formatted based on the Smart Patient Record format definition developed in WP2. T4.3 of WP4 is closely related to WP2 and WP5. The data technology developed in T4.3 will be applied to verify quality of fabricated medical data and its usage for data analytics and authentication tools of WP2 and WP5. In addition, the output of T4.4 will be used by the WP2 of the project.

1.3 Structure of this Document

This document is structured as follows: *Chapter 2* describes IBM's Data Fabrication Technology, its usage for masking of the project real medical data and fabrication of synthetic data. The chapter also provides samples of initial versions of the fabricated use-case data. *Chapter 3* describes the methodology that is used for verification of the fabricated data quality and its usage for development and testing of the project advanced data analytics and user authentication tools. *Chapter 4* provides a description of the semantic- and privacy-preserving encryption methodology that is used to enable and facilitate the application of the project advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations. *Chapter 5* concludes the deliverable.

2 Data Masking and Synthetic Data Fabrication

2.1 Introduction to DFP

IBM's Data Fabrication Platform (DFP) [1][2] is a web based central platform for generating high-quality data for testing, development, and training. The platform provides a consistent and organizational wide methodology for creating test data. The methodology used is termed "rule guided fabrication".

The primary DFP use case for fabricating synthetic data contains two actors: a user (initiator) and Database/File (participator). This use case includes two sub-use cases: data requirements modelling and data generation. The data requirements use case includes three sub-use cases: resources and structure definitions, constraint rules definitions and fabrication configuration definitions. The data structure for databases (schema, tables, columns, etc.) is automatically imported, however structural hierarchy of data elements (structs, arrays, tables, fields, types) need to be manually defined by the user. The constraint rules are required to construct a model of the data and thus enable creation of meaningful realistic data vales. Input and output resources are standard relational databases (e.g., DB2, Oracle, PostgreSQL, SQLite), standard file formats (e.g., Flat file, XLS, CSV, XML, JSON) and streaming via MQTT protocol.

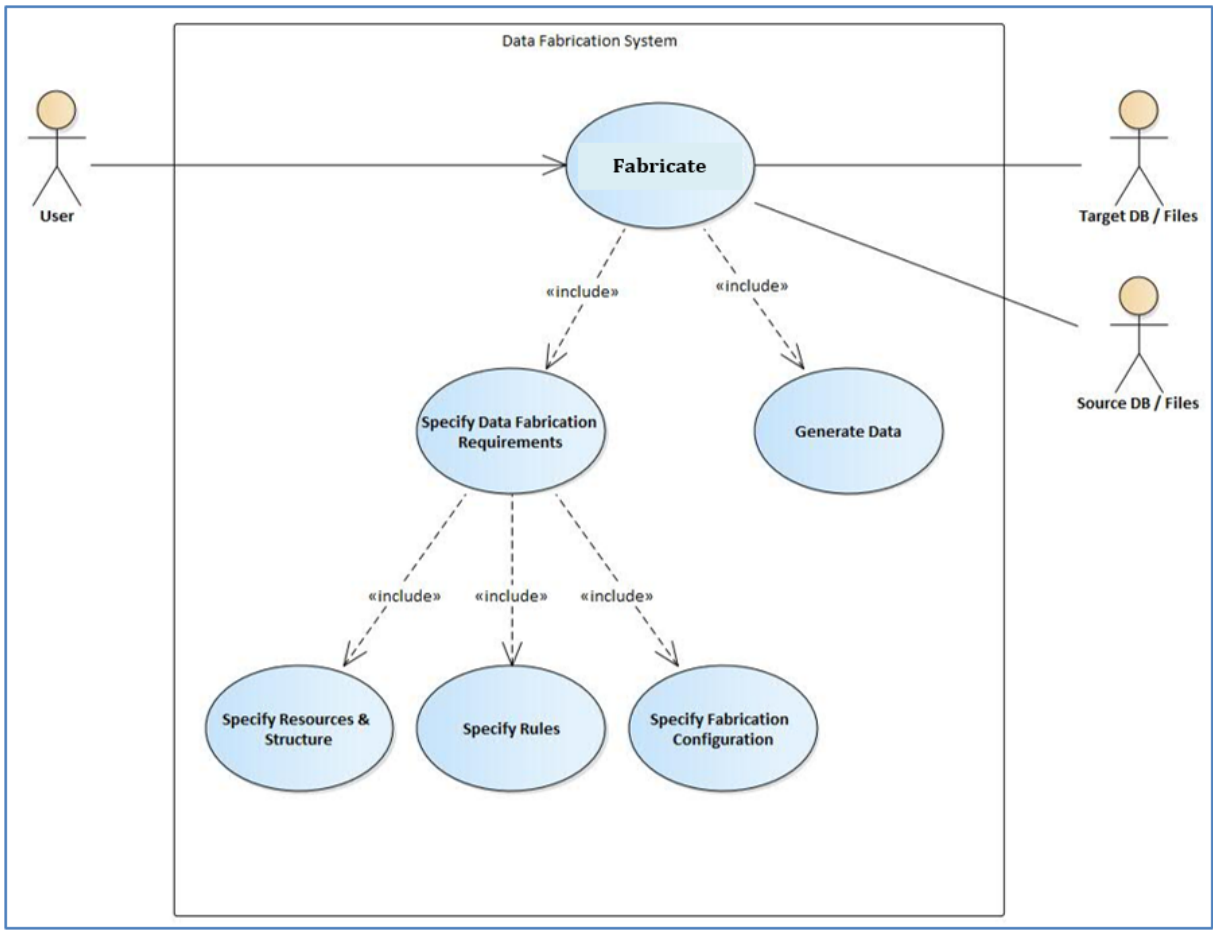


Figure 1 UML use case diagram for the DFP

In rule guided fabrication, the database logic is extracted automatically and is augmented by application logic and testing logic modelled by the user.

The application logic and the testing logic can be modelled using rules that the platform provides. The platform is also extendible and new rule types can be added by users and automatically integrated into the platform in an organization-wide manner.

Once the user requests the generation of a certain amount of data into a set of test databases, the platform internally ensures that the generated data satisfies the modelled rules as well as the internal databases consistency requirements.

The platform can generate data from scratch, inflate existing databases, move existing data, and transform data from previously existing resources, such as old test databases or even production data. The platform provides a comprehensive and hybrid solution that can create a mixture of synthetic and real data according to user requirements.

IBM Data Fabrication Platform uses a proprietary Constraint Satisfaction Programming solver (CSP) [3-24] that has been used for verifying IBM hardware systems for over a decade. The solver finds a solution for all the requirements in a data fabrication task. The solver is capable of handling very complex problems in a timely manner.

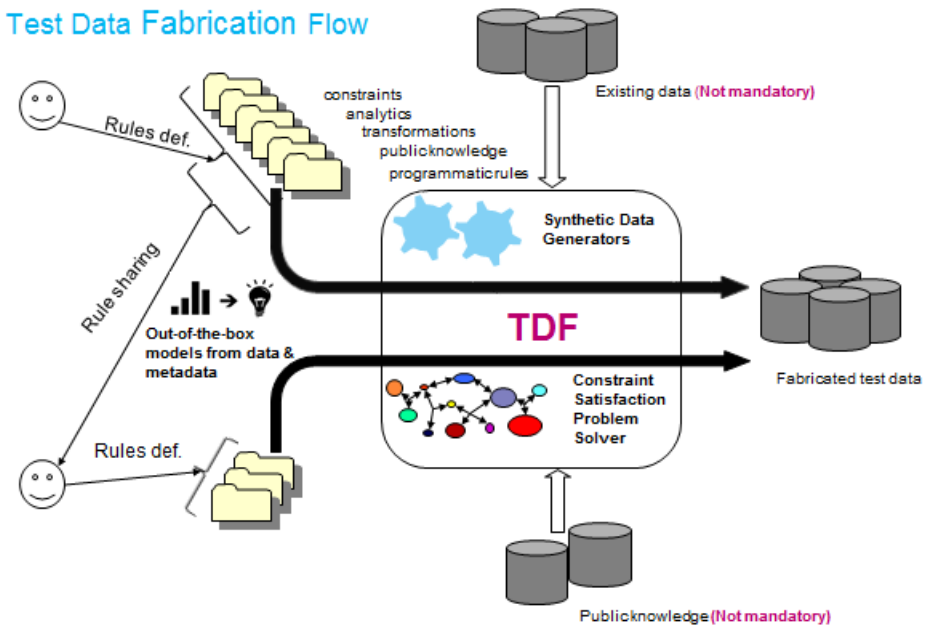


Figure 2 Data Fabrication Platform (DFP) flow

2.2 Synthetic Data Fabrication using CSP

To overcome the shortcomings of existing data generation techniques, DFP uses a solution that generates data using a Constraint Satisfaction Problem (CPS) solver. This methodology is generic and flexible for various types of use cases yet also very safe, as all user constraints must be satisfied. The solution does not require access to real or masked data, or to historic actual queries, which all might involve some violation of privacy regulations. Data generation

can be constrained directly by the users. These constraints can direct the generation towards desired testing objectives or to realistic database statistics and query behaviour.

The platform uses the database schema or the file hierarchical structure, the user requirements via variables and constraints and a fabrication configuration specifying which rules to use and where to write the generated data into. The constraint-problem is solved, and the solution is used to construct the records needed to populate the database or file.

The fabrication process is described next. A user has provided a data project which contains the structure of the data, the constraint rules and the fabrication configuration. In order to construct a constraint satisfaction problem for the solver, the platform analyses the table metadata and gets table columns' data type and other properties, e.g., referential integrity constraints. The platform selects a sub-set of the relevant rules and tables using the fabrication configuration. In addition, relevant parent tables may be added due to referential integrity dependencies. Additional default rules are sometimes required as well (PK, Unique Column, string and binary column widths, etc.).

This information is used for the construction of a database table dependency graph. For each table in that graph, starting at root nodes, structural record dependencies are built recursively.

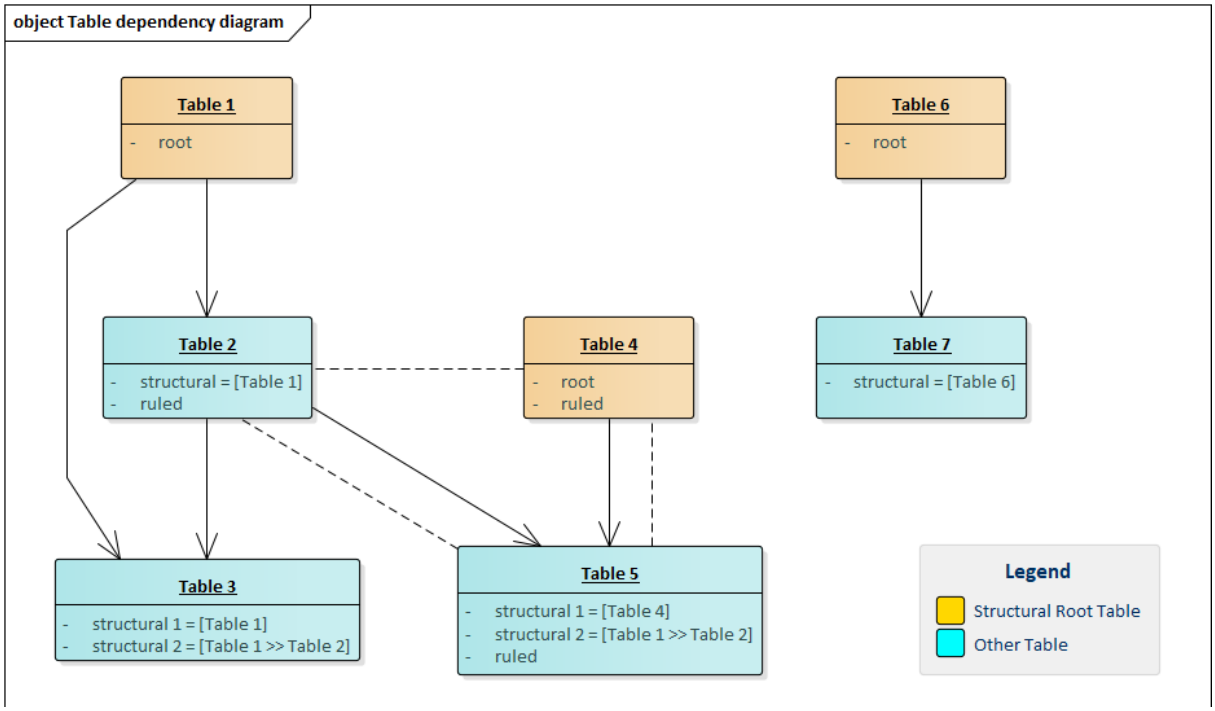


Figure 3 Table dependency diagram

Once the above graphs are built, the fabrication pattern is computed where each target table record is assigned to one of the following fabrication modes: 'New', 'Reuse' or 'Other'.

Given the patterns, the graph and the rules, a CSP problem can be created. The CSP has a language used to model a real-world problem. A problem consists of variables and rules. The solution is an assignment of one value for each variable where all the rules are satisfied. The solver finds a random solution. Each time it solves a problem, a new solution is produced. The user does not specify how to solve the problem, but rather focuses on what to solve using the specified rules for a legal solution.

Each table is translated into a CSP variable struct or a CSP vector of variable structs, according to the record type, each table column is translated into a CSP variable. Each of the rules is added to the problem as a CSP constraint. A simple example of the CSP problem structure can be seen in the figure below. The CSP problem defined here is the famous eight queens puzzle.

```
variable integer arr[8];

constraint queens: forEach (q, 0, sizeOf(arr)-1, 0 <= arr[q] <= 7);
constraint sameRow: allDiff (q, 0, sizeOf(arr)-1, arr[q]);
constraint sameDiag: forEach (i, 0, sizeOf(arr)-1, \\
                    forEach (j, 0, sizeOf(arr)-1, ((i!=j) -> (abs(i-j) != abs(arr[i]-arr[j])) )) );
```

Figure 4 A CSP example

Finally, the problem is submitted to the solver. The solution is recursively parsed starting the iteration root following the topology of (vectored) record variables. In the case of database project, an SQL insert statement for all table records is created and that SQL insert statement is submitted to the DB for execution. If streaming option is enabled and properly configured, the solution is converted to the relevant format and sent to the messaging broker specified in the configuration. In the case of file projects, the solver result is converted into the relevant file format.

2.3 Fabrication of Synthetic Healthcare Data

IBM Data Fabrication Platform is used in the SERUMS project to fabricate synthetic healthcare data for all the project real-world use cases. Below is the description of the initial medical data that have been created by the DFP tool to enable the development of the SERUMS Patient-Centric Healthcare Systems. All the data was inserted into the PostgreSQL databases.

2.3.1 Zuyderland Medisch Centrum Smart Health Centre (ZMC) Use Case

For this use case we have fabricated initial version of patients' personal medical data based on the data format and description provided by the ZMC hospitals.

This synthetic use case data consists of two database tables including patients' personal data (id, name, surname, etc.) and daily data extracted from their wearable devices (walking time, sitting time, cycling time, etc.). Multiple single-column and cross-column fabrication rules define legal value domains and relationships for each table column. Below is a sample data fabricated for the wearable devices table:

patient_nr	day_nr	time total	time passive	time active	time sit	time stand	time walk
1	1	29287	9919	19368	7581	2338	19368
2	1	49328	30105	19223	30104	1	19223
3	1	69855	57994	11861	53103	4891	11861
4	1	62178	54860	7318	43671	11189	7318
5	1	62214	51522	10692	41985	9537	5975
6	1	39775	26635	13140	25396	1239	13140
7	1	66680	54885	11795	52056	2829	11795

8	1	71115	58274	12841	49278	8996	12841
9	1	71165	62930	8235	53069	9861	7289
10	1	67850	55551	12299	51786	3765	12299
11	1	68621	54231	14390	46300	7931	14390
12	1	71340	64181	7159	45048	19133	7159
13	1	22074	175	21899	89	86	21899
14	1	28641	9418	19223	4238	5180	19183
15	1	47013	35714	11299	29726	5988	4598
16	1	42524	32250	10274	22177	10073	10274
17	1	62088	51076	11012	40834	10242	6978
18	1	63508	55835	7673	41232	14603	7058
19	1	34310	26697	7613	21204	5493	7613
1	2	67276	42243	25033	35106	7137	14658
2	2	37144	17745	19399	10982	6763	19399
3	2	66880	59108	7772	59105	3	7772
4	2	65299	45795	19504	42055	3740	19504

2.3.2 Hospital Clinic of Barcelona Smart Platform (FCRB) Use Case

This use case includes integrated healthcare data from the Catalan patient healthcare ecosystem, combining data generated inside the hospital with data created outside of it.

The synthetic use case data consists of patients' personal data (id, name, surname, etc.), medical episode data (episode identifier, medical centre id, category of the treatment, episode status, etc.), medication and prescription data, diagnostic data, professional identification data, and so on. It is currently organized as nine database tables with a complex topology. A general table diagram is presented in Figure 5 below.

Below is a sample data from the Episode description table:

EINRI	**FALNR**	**FALAR**	**PATNR**	**BEKAT**	**EINZG**	**STATU**	**KRZAN**	**ENDDT**
BCL	1E+09	1	1	B2023G	1	P	X	15/08/2019
BCL	1E+09	2	2	BCD015	2	I	X	02/09/2019
BCL	1E+09	1	3	BCO067	3	E		12/08/2019
BCL	1E+09	2	4	BC2101	4	E		08/09/2019
BCL	1E+09	2	5	B1001A	5	E	X	24/08/2019
BCL	1E+09	3	6	BCO086	6	E		12/06/2019

BCL	1E+09	1	7	BCD040	7	E	X	15/08/2019
BCL	1E+09	1	8	B5011P	8	P	X	12/08/2019
BCL	1E+09	2	9	BCO148	9	P		12/08/2019
BCL	1E+09	3	10	BSALA1	10	P	X	08/09/2019
BCL	1E+09	1	11	BC102C	11	E		21/05/2019
BCL	1E+09	1	12	2014G	12	P	X	09/09/2019
BCL	1E+09	3	13	10502	13	P	X	07/04/2019
BCL	1E+09	2	14	BCV009	14	I	X	23/03/2019
BCL	1E+09	3	15	B1001C	15	I	X	01/09/2019
BCL	1E+09	1	16	BCO082	16	E	X	29/08/2019
BCL	1E+09	3	17	BCO034	17	E	X	16/07/2019

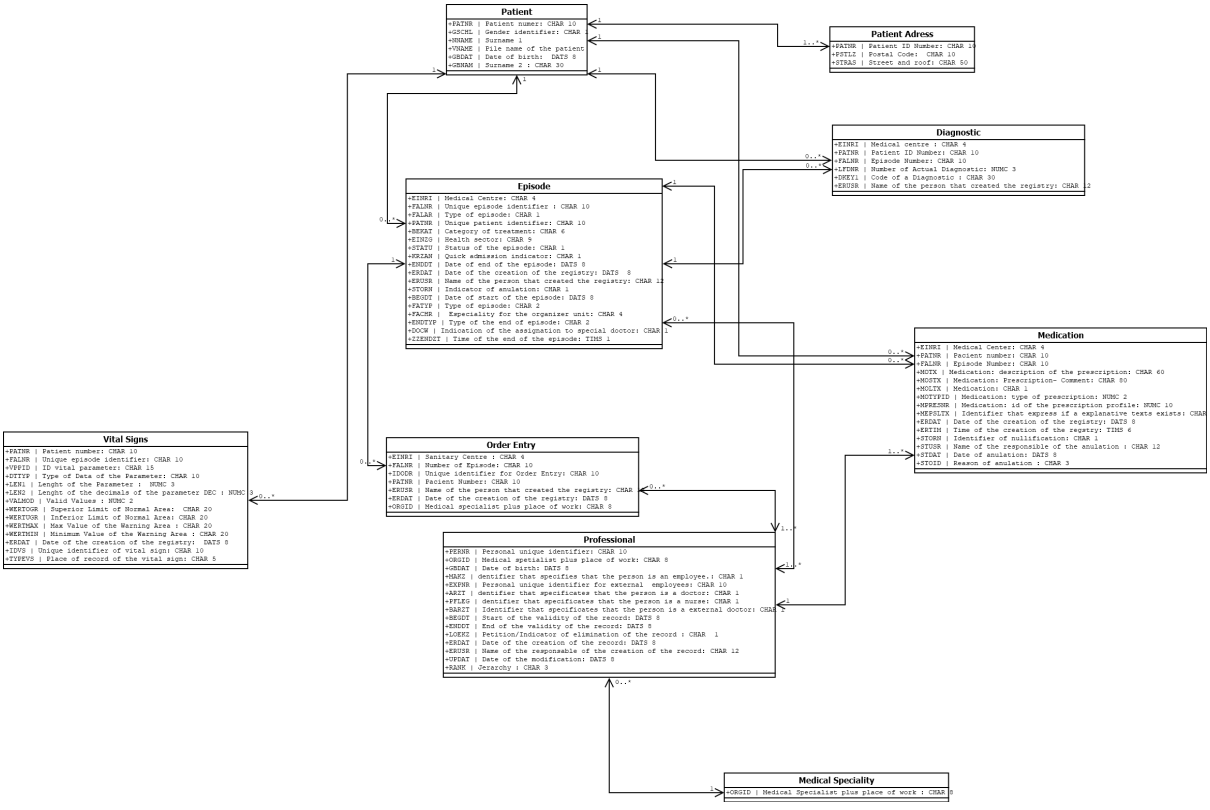


Figure 5 FCRB Generic Table Diagram

2.3.3 Personal Cancer Treatment (USTAN) Use Case

This use case includes healthcare data from the Edinburgh Cancer Care treatment centre. It is organized as a collection of six interconnected tables including personal data, treatment data, patient's clinical and healthcare conditions. Below is a sample of fabricated data from the Chemocare Treatment table.

appointment_date	last_toxicity_date	tumour_group	age_at_diagnosis	height	weight
03/07/2019	20/09/2019	Other	30	1.95	63.2
04/08/2019	19/09/2019	Breast	35.2	1.78	72.7
14/08/2019	17/09/2019	Lung and Chest	46.8	1.57	70.4
31/08/2019	24/09/2019	Breast	88.8	1.85	69.2
16/09/2019	16/09/2019	Lung and Chest	26	1.4	57.4
16/08/2019	07/09/2019	Other	20.4	1.43	81.2
21/08/2019	17/09/2019	Breast	64.2	1.66	78.9
18/06/2019	23/09/2019	Other	30.2	1.4	61.8
30/08/2019	19/09/2019	Other	33.6	1.04	79.6
21/07/2019	14/09/2019	Breast	24.5	1.91	83
14/09/2019	19/09/2019	Lung and Chest	31.9	1.73	71.4
03/06/2019	21/09/2019	GI Lower	22.4	1.03	72.3
24/08/2019	12/09/2019	Other	29.1	1.76	63.8
24/06/2019	14/09/2019	Breast	58.3	1.58	68.2
19/09/2019	19/09/2019	Other	78.2	1.71	80.5
24/08/2019	22/09/2019	Other	37.5	1.87	88.5
26/06/2019	13/09/2019	Other	48.8	1.66	78
19/07/2019	18/09/2019	Breast	30.6	1.27	82.7
05/06/2019	17/09/2019	Breast	69.7	1.25	59.2
03/07/2019	15/09/2019	Lung and Chest	30	1.95	63.2

2.4 Data Masking

Data masking is a well-known method of creating a structurally similar but inauthentic version of an organization's data that can be used for purposes such as software testing, software development and user training. The purpose is to protect the actual personal or sensitive data while having a functional substitute for occasions when the real data is not required. In data masking, the format of data remains the same, only the values are changed. The data may be altered in several ways, including encryption, character shuffling, and character or word substitution.

It was the SERUMS consortium decision that most of the data used for the development and testing of the SERUMS data analytics, user authentication technologies and its patient-centric

healthcare system will be synthetic data fabricated by IBM's Data Fabrication Technology described in Section 2.1 above. Moreover, usage of synthetic realistic data solves a known weakness of the data masking approach – its reversibility and a need for the real data access. In case synthetic fabricated data will not be sufficient or "good enough" for the development and testing requirements of the project, we will consider applying the same IBM's DFP tool to produce masked data from the project use cases real data.

3 Verification of Fabricated Data Quality

Since the fabricated data is used for testing, verification, and validation of many other parts of the SERUMS project, the quality of the fabricated data is crucial. To this end we have developed machine learning (ML) techniques to verify the quality of the data fabrication, and to provide feedback on how to improve the data fabrication within the scope of the SERUMS project.

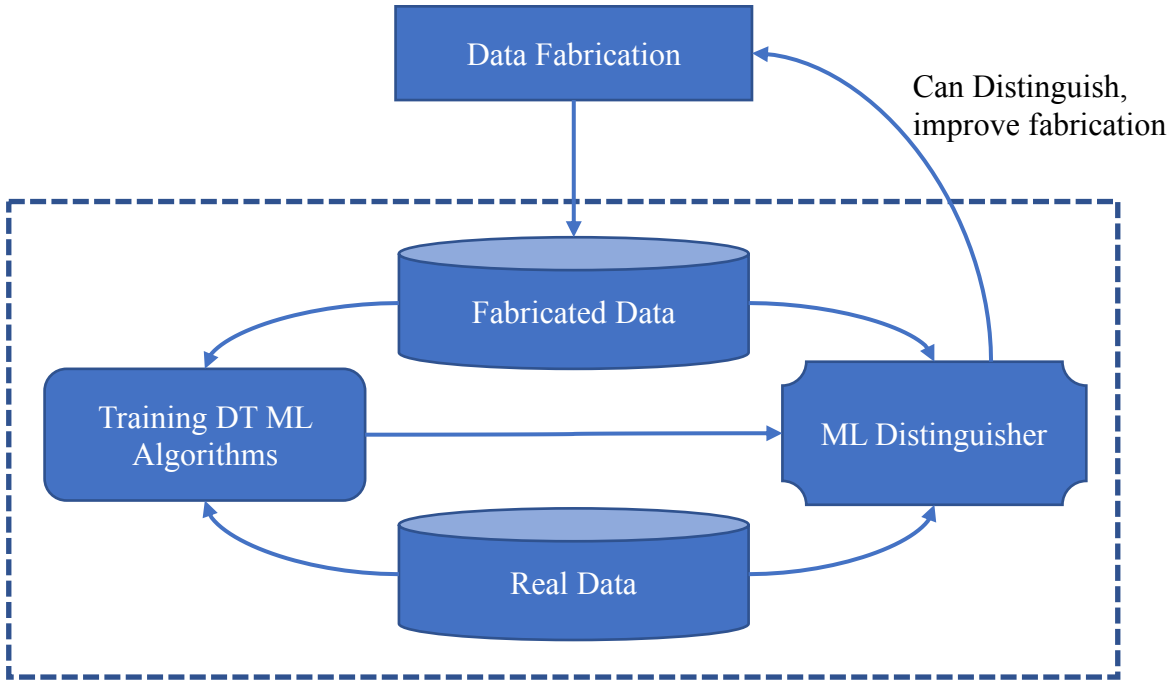
The data fabrication techniques (as described above) ensure that the fabricated data is correct. In combination with expert knowledge from our partners USTAN, ZMC, and FCRB this can be guided to produce data that appears both correct formally and reasonable medically. However, this does not ensure that the fabricated data matches the greater data patterns and less obvious intrinsic dependencies that may exist.

To address this challenge, we have developed several ML algorithms that can anonymously be trained to (attempt to) distinguish the fabricated data from the real data. Here we have chosen to use decision-tree (DT) based ML algorithms, since examination of the DTs can provide feedback on where the data fabrication is most distinct from real data.

These algorithms have been designed to run over data sets and discard all identifying features, revealing only the feature index that provides the decision on how to distinguish fabricated and real data. The index map to the feature name is kept and so we can identify which features of the fabricated data are used to distinguish.
This information can in turn be used to improve the data fabrication, by identifying where the existing constraints or probabilities are insufficient to produce "realistic" data.

Once this information has been used to improve the data fabrication, the ML algorithms can be run again comparing the real data with newly generated fabricated data, and the above process repeated. This can be done until the ML algorithms cannot effectively distinguish fabricated and real data, thus indicating that the data fabrication is of suitable quality.

An overview of this process is shown in the figure below.



Initial experiments have been conducted using the above process on fabricated and real data based on information from USTAN. These included more than 10 experiments, each iterating over 30 or more iterations (total more than 300 ML distinguishers trained and tested). These have identified several places where the data fabrication can be improved, as well as aided in improving the algorithms and process itself.

The results of these experiments have so far demonstrated the effectiveness of the ML distinguishers in both: distinguishing fabricated from real data, and identifying which features are significant in this distinguishing. Unfortunately for privacy reasons no detail of these results can be reported at this time.

4 Semantic-preserving Data Encryption

In recent years neural networks have been a prominent classification tool many times outperforming all other conventional machine learning algorithms. This started in the ImageNet project where a large visual database designed for use in visual object recognition software served as a benchmark for comparison. In the 2012 competition a new technique using Deep Neural Network won the competition by a margin. Since then the use of neural network has been growing steadily, with increased accuracy, as well as expanding to other classification scenarios and use cases including the healthcare domain. In healthcare as in many other fields, such as finance, the data may not be easily exposed due to different privacy regulations or lack in local resources. This in turn set the need for a way to run these neural network analytics on encrypted data.

Nowadays, several commercial services that perform detection of various types of diseases or medical analysis interpretation can be found in the market. These services can collect patients' data via dedicated wearable devices, analyse them to detect “anomalies” and report the results to a healthcare professional, who creates a report. Employing machine learning techniques (e.g. deep learning) is a good approach for improving the performance of automated medical data analysis. Unfortunately, there are limitations to this approach. Analysing long streams of healthcare data for many patients may be difficult to be handled on premises, where the potential lack of computational resources would limit the performance. To overcome this issue, one could acquire healthcare data on premises and outsource them to an external environment (with more resources), where the data analysis and potential illness detection would be performed. Nevertheless, moving from a trusted environment to an untrusted one would endanger the protection of personal data. Hence, it becomes essential to protect data before outsourcing them to the untrusted environment, e.g., via the use of advanced cryptographic techniques.

In order to evaluate the suitability and efficiency of the advanced cryptographic techniques researchers from IBM have developed a privacy-preserving arrhythmia classification tool based on neural networks. The team has started from building a small NN model, which is compatible with the use of Fully Homomorphic Encryption (FHE) for the classification of arrhythmia.

The size of the model is optimized to efficiently support the underlying cryptographic techniques. The input vector (remains unchanged and) is of size 180. The network consists of 2 fully-connected layers (40 hidden neurons in total) and 1 activation layer that uses x^2 as the activation function. The output vector is of size 16. This model achieves 96:24% accuracy.

Once this NN model is designed, it is applied it on the encrypted inputs. The proposed and implemented Hybrid solution follows the GAZELLE approach presented in [25]. In particular, the team has implemented the linear operations such as vector/matrix multiplication using FHE and the non-linear ones such as operations in activation layer by using Multi Party Computation (MPC). The implementation of the homomorphic encryption is based on the HElib.

Initial evaluation results of applying this approach for arrhythmia classification show that the approach is practically applicable. In the future we plan to try to apply the same homomorphic encryption technology and approach to analyse the SERUMS use case data.

5 Conclusions

The aim of this deliverable D4.1 is to report and describe the design and development of initial versions of the project data masking, data fabrication, data quality verification and semantic-preserving data encryption technologies. All these technologies are used to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart health-care system and to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

First, the document describes IBM's Data Fabrication Technology and its rule-based data fabrication approach to produce synthetic realistic medical data that is used for development and testing of all the project technologies. Initial synthetic data samples fabricated based on the data format and data characteristics defined and provided by the project use-case data owners is also presented in the document. Further, the document describes our approach to estimating the quality of fabricated synthetic data to ensure that all data analytics and user authentication tools developed by Serums consortium will be fully applicable for real medical data in the future. The document also describes our approach to semantic- and privacy-preserving data encryption to be able to apply the advanced data analytics and machine-learning algorithms for analyzing encrypted personal data.

This document is the first deliverable of Work Package 4: "Secure and Privacy-Preserving Data Communication". Deliverables D4.2 and D4.3 will describe more advanced versions of all the above technologies and their application for the development of the Serums Smart Health Centre system.

References

[1] “Create high-quality test data while minimizing the risks of using sensitive production data.” *IBM InfoSphere Optim Test Data Fabrication*, IBM, 2017, https://www.ibm.com/il-en/marketplace/infosphere-optim-test-data-fabrication.

[2] “Test Data Fabrication.” *Security and Data Fabrication*, IBM Research, 2011, https://www.research.ibm.com/haifa/dept/vst/eqt_tdf.shtml.

[3] “Constraint Satisfaction.” IBM Haifa Research, IBM, 2002, https://www.research.ibm.com/haifa/dept/vst/csp.shtml.

[4] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors (2007), "Optimatch: Applying Constraint Programming to Workforce Management of Highly-skilled Employees", International Journal of Services Operations and Informatics (IJSOI), Vol 3, No. 3/4, pp. 258 - 270.

[5] Y. Naveh, Y. Richter, Y. Altshuler, D. Gresh, and D. Connors (2007), "Workforce Optimization: Identification and Assignment of Professional Workers Using Constraint Programming", IBM J. R&D.

[6] Y. Naveh, M. Rimon, I. Jaeger, Y. Katz, M. Vinov, E. Marcus, and G. Shurek (2006), "Constraint-Based Random Stimuli Generation for Hardware Verification", AI magazine Vol 28 Number 3.

[7] E. Bin, R. Emek, G. Shurek, and A. Ziv (2002). "Using a constraint satisfaction formulation and solution techniques for random test program generation", IBM Systems Journal, 2002.

[8] Merav Aharoni, Odellia Boni, Ari Freund, Lidor Goren, Wesam Ibraheem, Tamir Segev (2015), "Rectangle Placement for VLSI Testing", CPAIOR 2015: 18-30

[9] O. Boni, F. Fournier, N. Mashkif, Y. Naveh, A. Sela, U. Shani, Z. Lando, A. Modai (2012) "Applying Constraint Programming to Incorporate Engineering Methodologies into the Design Process of Complex Systems" Proceedings of the Twenty-Fourth Conference on Innovative Applications of Artificial Intelligence, Toronto, Ontario, Canada. AAAI 2012.

[10] Y. Ben-Haim, A. Ivrii, O. Margalit and A. Matsliah (2012) "Perfect Hashing and CNF Encodings of Cardinality Constraints", SAT 2012, Trento, Italy.

[11] E. Bin, O. Biran, O. Boni, E. Hadad, E. K. Kolodner, Y. Moatti, D. H. Lorenz (2011), "Guaranteeing High Availability Goals for Virtual Machine Placement", ICDCS 2011.

[12] Jeonghee Shin, John A Darringer, Guojie Luo, Merav Aharoni, Alexey Y Lvov, G Nam, Michael B Healy (2011), "Floorplanning challenges in early chip planning", SOCC Conference, 2011 IEEE International, pp. 388—393

[13] Y. Naveh (2010). "The Big Deal, Applying Constraint Satisfaction Technologies Where it Makes the Difference". Proceedings of the Thirteenth International Conference on Theory and Applications of Satisfiability Testing (SAT'10).

[14] S. Asaf, H. Eran, Y. Richter, D. P Connors, D. L. Gresh, J. Ortega, M. J. Mcinnis (2010). "Applying Constraint Programming to Identification and Assignment of Service Professionals". Accepted for presentation in The 16th International Conference on Principles and Practice of Constraint Programming (CP2010). The paper received the Best Application Paper Award.

[15] B. Dubrov, H. Eran, A. Freund, E. F. Mark, S. Ramji, and T. A. Schell, (2009). "Pin Assignment Using Stochastic Local Search Constraint Programming" in Proceedings of the 15th International Conference on Priniciples and Practice of Constraint Programming (CP'09), Edited by Ian P. Gent, pp 35-49.

[16] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors (2007), "Optimatch: Applying Constraint Programming to Workforce Management of Highly-skilled Employees", IEEE/INFORMS International Conference on Service Operations and Logistics, and Informatics (SOLI), Philadelphia, pp. 173-178.

[17] S. Sabato and Y. Naveh (2007), "Preprocessing Expression-based Constraint Satisfaction Problems for Stochastic Local Search", Proceedings of The Fourth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR).

[18] Y. Naveh, M. Rimon, I. Jaeger, Y. Katz, M. Vinov, E. Marcus, and G. Shurek (2006), "Constraint-Based Random Stimuli Generation for Hardware Verification", IAAI 2006.

[19] Y. Richter, A. Freund, and Y. Naveh (2006), "Generalizing AllDifferent: The SomeDifferent constraint", Proceedings of the 12 International Conference on Principles and Practice of Constraint Programming - CP 2006, Lecture Notes in Computer Science, Volume 4204, pages 468-483.

[20] Y. Naveh and R. Emek (2006). "Random stimuli generation for functional hardware verification as a CP application - a demo", IAAI 2006.

[21] Y. Naveh (2005). "Stochastic solver for constraint satisfaction problems with learning of high-level characteristics of the problem topography" CP 2005

[22] F. Geller and M. Veksler (2005), "Assumption-based pruning in conditional CSP", in van Beek, P., ed., CP, "Principles and Practice of Constraint Programming - CP 2005" of Lecture Notes in Computer Science (3709), 241-255 Springer.

[23] R. Dechter, K. Kask, E. Bin, and R. Emek (2002). "Generating random solutions for constraint satisfaction problems", AAAI 2002.

[24] D. Lewin, L. Fournier, M. Levinger, E. Roytman, G. Shurek (1995). "Constraint Satisfaction for Test Program Generation", Internat. Phoenix Conf. on Computers and Communications, March 1995.

[25] Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.: Gazelle: A low latency framework for secure neural network inference. arXiv preprint arXiv:1801.05507 (2018).