Health Market Inquiry

Health Market Inquiry

Promoting Healthy Competition

competitioncommission south africa

Standard Operating Procedure (SOP) for Data De-Identification

1. PURPOSE & SCOPE

The purpose of this Standard Operating Procedure (SOP) is to outline the preparatory and procedural steps to be followed in the de-identification of personal identifiers in respect of data submitted to the Health Market Inquiry (HMI), by stakeholders. The aim of the SOP is to advise stakeholders of the data collection process to ensure cooperation with the HMI. This document should be read with, and applied in accordance with, the **DE-IDENTIFICATION OF PERSONAL DATA** document published by the HMI on 1 June 2015 as well as related HMI policies.

Parties who wish to consult the HMI with specific issues relating to this document may contact the Inquiry Director.

2. DATA DE-IDENTIFICATION PREPARATION STEPS

2.1. The format of the data required by the HMI will vary from stakeholder to stakeholder. Stakeholders are urged to kindly contact the HMI for the applicable *Data File Specification Document (DFS)*.
2.2. The de-identification of data refers to the de-identification of address information, as well as personal identifiers such as name, surname, date of birth, identity numbers or any other data fields which could potentially be used to identify individuals.
2.3. All stakeholders will use the same de-identification algorithm for both address data de-identification and personal identifiers de-identification to ensure consistency in the HMI de-identification process.
2.4. Please read the attached flow diagrams in conjunction with the following steps (contained in this SOP) in preparation of the de-identification process. (**Appendix A** illustrates the end-to-end data process. **Appendix B** provides detail of the de-identification process).
2.5. The data to be de-identified may either be submitted in full to the HMI, or separated into three parts for preparation; *address data tables*, *personal identifier tables*, and the *claims/billing data tables*.
2.6. Where stakeholders require the HMI to perform de-identification on their entire data sets, the HMI will proceed by completing the protocol as set out in Step 4 of this SOP.
2.7. In instances where the data is separated by the stakeholder for de-identification, the following steps must be followed:
    2.7.1. Address data must be supplied to the HMI in order for the HMI to de-identify it. The HMI will then return the address data together with Enumerator Area Codes (EA Codes) to the stakeholder. The stakeholder will insert the appropriate reference codes as per the DFS, remove the identified address data, and submit the de-identified address data table to the HMI. In this regard case the HMI may engage with the stakeholder to clarify address data issues, should any exist.
    2.7.2. Personal Identifiers can either be de-identified by the stakeholder directly on the HMI de-identification server, whereby the stakeholder will split the Output file from the Input file and add applicable reference codes to the Output file as per DFS. This Output file will be submitted to the HMI. The HMI will provide stakeholders with individual usernames, passwords and a server address to proceed with this step.
    2.7.3. In instances where the stakeholder elects to de-identify personal identifiers at the designated premises of the HMI service provider, the protocol in Step 4 of this SOP will be followed.
    2.7.4. Claims/billing data is to be supplied to the HMI containing applicable reference codes, as per DFS.
    2.7.5. It is the responsibility of the stakeholder to ensure that referential integrity is maintained within all datasets as per DFS.
2.8. Should the stakeholder be aware of potential data issues, concerns or inconsistencies that might occur in the datasets, then the stakeholder is required to furnish the HMI with written details regarding such issues, concerns or inconsistencies, as well as possible resolutions for these matters during data submission.

3. OUTCOME OF THE DATA PREPARATION STEPS

3.1. Stakeholders requiring an explanation of the de-identification process may engage directly with the HMI.
3.2. Should a stakeholder require data encryption during the de-identification process, it may do so by employing the HMI's encryption algorithm. In such case:

3.2.1. The HMI will provide stakeholders with individual usernames, passwords and a server address to proceed with this action.
3.2.2. The Stakeholder will create a unique key for the data encryption process.
3.2.3. Output de-identified data files will be encrypted using the encryption process and encryption key.
3.2.4. The HMI will receive the encryption key from the stakeholder representative once encryption of data files is completed.

4. DATA DE-IDENTIFICATION PROTOCOL

The de-identification protocol (Step 4) applies to the *all three de-identification option*s: (1) de-identification by the designated HMI service provider; (2) self-service via the secure HMI server connection or; (3) on-site at the HMI service provider's premises.

4.1. The stakeholder and the HMI will use the Sign-off Document to record all steps followed in the de-identification process for record purposes.
4.2. De-identification takes place at the agreed time and place and utilises the HMI's de-identification server. The de-identification server contains the following applications and data: a) batch geo-coder, b) hashing application for personal de-identification purposes, c) geo-address de-id application, d) encryption application, e) data validation tool, f) MD5 Checksum Tool, g) census data and h) South African National address dictionary.
4.3. The status of the de-identification server must be checked to ensure that it is in an appropriate condition *before* loading of data commences.
4.4. The HMI loads the provided datasets to compare against the Data Quality Management Protocol (DQMP) and verifies adherence to the provided measurements. If the data quality checks fail, the data issues will be discussed with the HMI, and the stakeholder may thereafter be required to address any relevant data concerns and to re-submit the data.
4.5. After the stakeholder and HMI are satisfied with the data quality of the input file, the MD5 Checksum Tool is used to measure the INPUT file and the results are recorded on the Sign-Off Document.
4.6. The geo-coding step transforms the address data to the required GPS coordinates.
4.7. INPUT row lines and OUTPUT row lines are validated to ensure all records were processed.
4.8. The geo-address de-identification application transforms the GPS coordinates into an EA Code.
4.9. INPUT row lines and OUTPUT row lines are validated to ensure all records were processed during the address de-identification process.
4.10. The personal identifier data tables are de-identified by means of the "JOAAT Soft Hashing" algorithm. The OUTPUT file from this process includes the INPUT personal identifiers, as well as the OUTPUT de-identifiers.
4.11. The stakeholder and HMI validate that the line and header totals are consistent between the OUTPUT file and the original INPUT data file.
4.12. The HMI will conduct spot checks according to the DQMP on the INPUT/OUTPUT data file to ensure that correct de-identification has been achieved.
4.13. The INPUT data and the OUTPUT data are separated and the Stakeholder retains the INPUT data only.
4.14. The OUTPUT file is measured by the MD5 Checksum tool and the checksum is recorded on the Sign-Off Document.
4.15. The HMI receives the data files on the transfer file medium as supplied and marked by the stakeholder.
4.16. Both parties sign-off the received data file and transfer file medium.
4.17. All data created during the process is treated in accordance with the File Management Protocol.
4.18. The HMI finalises the Sign-Off Document.
4.19. The HMI receives the de-identified data and processes the receipt and data.
4.20. The HMI receives the de-identified data which processes and manages it as set out in **Appendix A**.

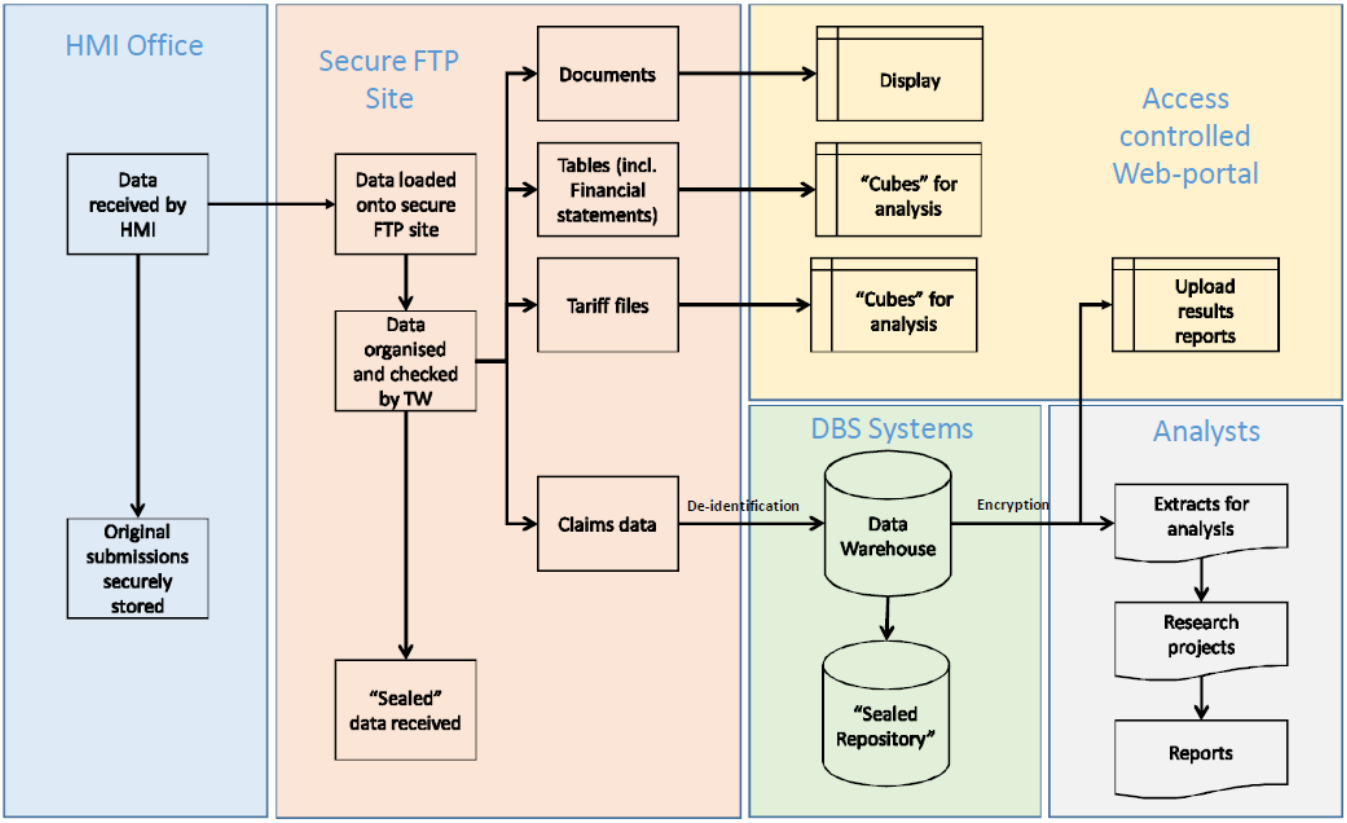
5. POST DATA QUERIES ON DE-IDENTIFIED DATA

5.1. The HMI will inform the Stakeholder if there are any queries relating to the submitted de-identified data.

5.2. If the stakeholder needs to relate its INPUT data to the OUTPUT data that is being queried; the stakeholder must arrange with the HMI for a data de-identification process.
5.3. This data de-identification process is performed in terms of this Sign-off Document.
5.4. The original Sign-off Document is used as control tool for the de-identification process to ensure all original recorded quality measurements correlate with new measurements taken during this process.
5.5. The HMI confirms that the provided INPUT data files are the same INPUT files as used in the original de-identification process. This is done by comparing file checksums with the MD5 Checksum tool against original measurements in the original Sign-Off Document.
5.6. OUTPUT files are generated from the INPUT files according to the same protocol as in Step 4 of this document.
5.7. The Stakeholder uses the INPUT and OUTPUT data files to clarify HMI queries.
5.8. All data created during the process is treated in accordance with the File Management Protocol.
5.9. The original Sign-Off Document is updated to record this process.

END OF DOCUMENT

APPENDIX A HMI DATA MANAGEMENT PROCESS FLOW



APPENDIX B DE-IDENTIFICATION PROCESS FLOW