# Report: GRT-INSPIRED SAPRIN Operational Database Data Ingestion

## Introduction

Data was transferred from all headquarters approved interviews on Survey Solutions with an interview key entry in the mapping spreadsheets into an empty copy (except for code values) of the SAPRIN standard operational database. In the process minor changes were made to the database structure to accommodate the specifics of the GRT-INSPIRED data.

Data was exported into tab-delimited text files using the Data Export features of Survey Solutions, harmonised and transformed and inserted into the database using [Pentaho](https://www.hitachivantara.com/en-anz/products/lumada-dataops/data-integration-analytics/pentaho-community-edition.html#:~:text=Pentaho%20Community%20Edition%20is%20an,9.3.0.0Version). The Pentaho code (jobs & transformations) used are available on [GitHub](https://github.com/SAPRIN-MRC/GRT-INSPIRED_ETL).

A key issue in the process was to identify the location (homestead, dwelling, apartment) at which the data associated with each interview was collected. Location identification information in the interviews were inconsistent and could not be relied upon to make this determination. It was decided to make use of ‘mapping files’ prepared by the GRT-INSPIRED spatial team (process undocumented) consisting of spreadsheets for each of the sites (Hillbrow, Melusi & Atteridgeville) listing each identified location and the interview key of the interview conducted at that location. The data ingestion process assumes that this linking is accurate and complete, specifically, the data in interviews not appearing (linked) in the mapping spreadsheets are not included in the data ingestion process.

## Locations

The reference base for the number of locations to be surveyed are all locations identified in the mapping spreadsheets. In the Atteridgeville mapping, backyard dwellings are identified as well. These have been ignored – specifically only the geocode “\*-001” have been considered (also all interview keys are associated with these geocodes). Table 1 summarises the outcomes with reference to the total set of locations by site.

* *Interview Linked*. Mapping spreadsheet contains an interview key associated with this location
* *Interview Saved*. The interview was in the headquarter approved interviews and an observation was successfully created in the database representing the interview. For this to happen, the interview must contain at least one contact with a valid contact date. The difference between this and the previous indicators, are either interviews not yet HQ-approved (28), or approved interviews without a valid contact, or an error in the interview key recorded on the spreadsheet (442).
* *Household Recorded*. At least one household was identified at the location. See interview section for more details.
* *Individuals Recorded*. One or more household members were recorded as belonging to the household/s at the location.

Table 1: Analysis of Mapped Locations

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Site | Locations | Interview Linked | % | Interview Saved | % | Household Recorded | % | Individuals Recorded | % |
| Atteridgeville | 9 879 | 8 688 | 88 | 8 559 | 87 | 5 107 | 52 | 5 064 | 51 |
| Hillbrow North | 10 280 | 10 280 | 100 | 10 086 | 98 | 6 902 | 67 | 6 668 | 65 |
| Hillbrow South | 6 997 | 6 997 | 100 | 6 701 | 96 | 3 273 | 47 | 2 478 | 35 |
| Melusi | 10 607 | 6 342 | 60 | 6 320 | 60 | 3 390 | 32 | 3 384 | 32 |
| Total | 37 763 | 32 307 | 86 | 31 666 | 84 | 18 672 | 49 | 17 594 | 47 |

## Interviews

An analysis of all interviews on Survey Solution (as of 3 Jan 2023) is summarise in Table 2.

Table 2: Analysis of Interviews

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  | No Household | | | | Household | |
| Site | Interviews | Not HQ Approved | HQ Approved | Not Mapped | Mapped | Not Saved | Saved | Non-contact | Refusal | No Dwelling | Undefined | Household | Individuals |
| Atteridgeville | 10 779 | - | 10 779 | 2 095 | 8 581 | 27 | 8 554 | 644 | 1 885 | 242 | 677 | 5 106 | 5 063 |
| Hillbrow | 17 873 | 9 | 17 864 | 796 | 16 817 | 30 | 16 787 | 5 041 | 217 | 406 | 948 | 10 175 | 9 146 |
| Melusi | 8 899 | 63 | 8 836 | 2 502 | 6 326 | 1 | 6 325 | 297 | 556 | 854 | 1 227 | 3 391 | 3 385 |
| Unknown | 833 | 33 | 800 | 790 | - | - | - | - | - | - | - | - | - |
| Total | 38 384 | 105 | 38 279 | 6 183 | 31 724 | 58 | 31 666 | 5 982 | 2 658 | 1 502 | 2 852 | 18 672 | 17 594 |

* *Interviews*. Total number of interviews on Survey Solutions.
* *Not/HQ approved*. Number of interviews with HQ approval or not.
* *Not/Mapped*. Number of interviews mapped to a location or not.
* *Not/Save*d. Number of mapped interviews saved to the database or not.
* *No Household*. No household captured in interview.
  + *Non*-contact. Contact attempt unsuccessful.
  + *Refusal*. Data collection refusal at dwelling
  + *No dwelling*. Broken down or non-functional dwelling.
  + *Undefined*. Unable to determine reason why no households have been identified.
* *Households*. Number of interviews with at least one household identified.
* *Individuals*. Number of interviews with at least one household member identified.

Table 3: Interview Indicators

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Site | Interview breakdown per site | Unmapped Interview % | Saved Interview % | Non-contact rate | Refusal rate | HH Enumeration rate | Individual Enumeration rate |
| Atteridgeville | 28% | 19% | 79% | 19% | 27% | 60% | 59% |
| Hillbrow | 47% | 4% | 94% | 38% | 2% | 61% | 54% |
| Melusi | 23% | 28% | 72% | 38% | 14% | 54% | 54% |
| Unknown | 2% | 95% |  |  |  |  |  |
| Total | 100% | 16% | 83% | 33% | 12% | 59% | 55% |

* The denominator for the Non-contact, Refusal, HH Enumeration and Individual Enumeration rates are mapped interviews.

## Households

An analysis of all households identified in mapped locations is presented in Table 4.

Table 4: Household Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Without Members | | | Indicators | |
| Site | Households | With Members | Participation Refused | No Consent | Undefined | Refusal rate | % Undefined |
| Atteridgeville | 5 209 | 5 125 | 55 | 1 | 28 | 1,1% | 0,5% |
| Hillbrow North | 9 154 | 7 832 | 247 | - | 1 075 | 2,7% | 11,7% |
| Hillbrow South | 4 580 | 2 924 | 755 | - | 901 | 16,5% | 19,7% |
| Melusi | 3 424 | 3 407 | 3 | 1 | 13 | 0,1% | 0,4% |
| Total | 22 367 | 19 288 | 1 060 | 2 | 2 017 | 4,7% | 9,0% |

* *Households*. Total number of households recorded on mapped interviews.
* *With Members*. Households with saved household members.
* *Without Members*. Households without members.
  + *Participation Refused*. Household participation refused (hhr\_permission = 2)
  + *No consent*. Did not consent to household questionnaire (hhr\_consent = 2)
  + *Undefined*. Not possible to identify why no household members have been captured.

## Individuals

An analysis of all individuals in mapped locations is presented in Table 5.

Table 5: Individual Analysis

|  |  |  |
| --- | --- | --- |
| Site | Individuals Recorded | Individuals Saved |
| Atteridgeville | 11 083 | 11 078 |
| Hillbrow North | 20 228 | 20 224 |
| Hillbrow South | 5 805 | 5 801 |
| Melusi | 6 323 | 6 322 |
| Total | 43 439 | 43 425 |

* *Individuals Recorded*. All individual recorded on interviews conducted at mapped locations
* *Individuals Saved*. Individuals successfully saved to the database

Table 6 contains an age and sex breakdown of the population by site. Site specific population pyramids are provided as well.

Table 6: Population Age and Sex breakdown

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Atteridgeville | | Hillbrow North | | Hillbrow South | | Melusi | | | Total | |
| AgeGroup | M | F | M | F | M | F | | M | F | M | F |
| 00-04 | 299 | 281 | 1 047 | 1 056 | 215 | 233 | | 352 | 299 | 1 913 | 1 869 |
| 05-09 | 278 | 251 | 821 | 783 | 144 | 161 | | 158 | 170 | 1 401 | 1 365 |
| 10-14 | 285 | 269 | 627 | 715 | 128 | 149 | | 137 | 145 | 1 177 | 1 278 |
| 15-19 | 323 | 364 | 560 | 681 | 310 | 453 | | 122 | 164 | 1 315 | 1 662 |
| 20-24 | 523 | 571 | 878 | 1 185 | 689 | 722 | | 291 | 365 | 2 381 | 2 843 |
| 25-29 | 580 | 648 | 1 007 | 1 367 | 309 | 408 | | 459 | 439 | 2 355 | 2 862 |
| 30-34 | 517 | 593 | 1 009 | 1 356 | 230 | 340 | | 499 | 437 | 2 255 | 2 726 |
| 35-39 | 525 | 500 | 1 084 | 1 295 | 229 | 248 | | 451 | 369 | 2 289 | 2 412 |
| 40-44 | 448 | 465 | 808 | 853 | 160 | 175 | | 310 | 241 | 1 726 | 1 734 |
| 45-49 | 371 | 381 | 692 | 622 | 118 | 107 | | 211 | 163 | 1 392 | 1 273 |
| 50-54 | 323 | 348 | 381 | 354 | 65 | 61 | | 132 | 100 | 901 | 863 |
| 55-59 | 250 | 312 | 243 | 192 | 32 | 39 | | 84 | 61 | 609 | 604 |
| 60-64 | 181 | 270 | 89 | 103 | 13 | 22 | | 41 | 45 | 324 | 440 |
| 65-69 | 153 | 181 | 56 | 51 | 7 | 12 | | 26 | 14 | 242 | 258 |
| 70-74 | 85 | 136 | 28 | 32 | 3 | 2 | | 13 | 8 | 129 | 178 |
| 75-79 | 48 | 114 | 14 | 18 | 1 | 3 | | 5 | 4 | 68 | 139 |
| 80-84 | 23 | 74 | 15 | 14 | 1 | 1 | | 1 | 1 | 40 | 90 |
| 85+ | 29 | 79 | 97 | 91 | 3 | 8 | | 2 | 3 | 131 | 181 |
| Total | 5 241 | 5 837 | 9 456 | 10 768 | 2 657 | 3 144 | | 3 294 | 3 028 | 20 648 | 22 777 |

|  |  |
| --- | --- |
|  |  |
|  |  |

## Discussion

The data ingestion process identified several issues that requires further attention. List of records with data integrity problems are contained in the folder “.\ETLOutput\DataIntegrity” for attention, unless otherwise noted.

1. **Inaccurate dates**. In several places the data collection form allows the entry of dates using the format “###\*‑\*\*‑\*\*”, as a placeholder for a so-called ‘variable precision date’, that uses ‘U’ characters to identify unknow parts of a date, e.g. decade, month or day. These entries need to be validated more thoroughly, because dates like “1021-UU-UU” or “2068-UU-UU” have been entered, so the first two digits, or year part, if it is fully numeric, should be validated for a plausible range. Dates such as “9999-99-99” are also common and should be disallowed. During the data ingestion process these dates were substituted with plausible derivations from the entered value (for details see ETL source code). The result of this issue, or just dates incorrectly recorded caused a few household-member records not to be saved, because of invalid or future dates of birth. Date relationships still to be validated are the age of mother at delivery to verify that these are plausible, and duration between different pregnancies of the same mother.
2. **Mapping file interview keys**. The mapping files play a crucial role in the data ingestion process, because only interviews corresponding to these keys are considered for ingestion into the database.
   1. **Invalid interview keys.** These are keys not associated with any interview in the Survey Solution data export. The reason for this is unknown but could be due to data entry errors into the mapping spreadsheet. There are 442 such records in Hillbrow North (147) and Hillbrow South (295). It is also possible that these interview keys refer to records in the Hillbrow BS Registration questionnaire, if this is the case a flag needs to be added to the mapping spreadsheet to clearly indicate this. These records are listed in the Statistics spreadsheet. There are also 28 interviews linked that are not yet HQ Approved (see Statistics spreadsheet).
   2. **Interviews not reflected in the mapping files.** There are 6 183 such interviews (See Table 2). These could be so-called “duplicate” interviews, but it seems a very large number. Given the fact that the total ingested population, 43 425, falls far below the SAPRIN target of 100 000, these interviews could represent unlinked interveiws that could increase the survey population, and should be investigated as a matter of urgency.
3. **Questionnaire design limitations.** There are a few design limitations in the questions on the questionnaire and comments have been made on the questionnaire design to point these out. It is probably too late to make any changes to the questionnaire now, but care should be taken not transfer these to the longitudinal surveillance questionnaire. The most important limitation is in the social grant questions, where the date of start of a grant and the beneficiary cannot be associated with a specific grant. Grants should be implemented as a roster. The maternity history section is also difficult to parse into the format required by SAPRIN. Reference should be made to the AHRI or DIMAMO questionnaires to address this.
4. **Missing households and household members.** There are 2 852 interviews where the reason for not recording a household (Table 2), cannot be determined, and 2 017 households (Table 4) where the reason for not recording any members can not be determined. Together with interviews not reflected in the mapping files, this is probably the biggest data collection problem and could represent a potentially large proportion of the population that that is not enumerated.
5. **Locations without a linked interview.** There are 5 456 locations (excluding backyard dwellings) without a linked interview. It is possible that these are not residential locations, or dwellings verified to be empty, but could also represent locations where the interview has not been linked, or a data collection visit has not taken place.

## Conclusion

It has been a labour intensive to develop the ingestion process, but the process is now fully automated and can be easily repeated by GRT-INSPIRED staff when data in Survey Solutions are updated, or new data captured.

The biggest issue is the fact that less than half the target SAPRIN nodal population has been enumerated. If we assume that the original boundaries of the node correctly demarcate a potential surveillance population of 100 000, then the reasons for the low enumerated population could be a combination of the following:

1. Community consent is low, restricting the number of individuals that can be successfully enumerated. With continued and effective community engagement, consent can be improved, and the enumerated population increased.
2. Not all residential locations have been surveyed, this could be due to known reasons, such as ‘captured’ buildings, or fieldwork related issues. These should be addressed to ensure as a complete as possible coverage of the surveillance area.
3. The mapping file issues identified above, will also result in missing interviews that should be included in the data ingestion process. This needs to be addressed urgently and the data ingestion process repeated.
4. Missing households and household members. The reason for this should be urgently investigated and addressed. This is a serious issue that could require revisits to locations with data collection problems.