Prepared for CNLG

# GACACA ARCHIVE PROJECT (GAP)

# PHASE V

## Data consolidation Report

## 1. Introduction

This report outlines the findings of the Metadata team that carried out the activities of handling data provided by CNLG, correcting apparent errors in the spreadsheets (flat file), modelizing and processing those data into a relational database management system (DBMS).

**Summary advantages of a relational database (DBMS) over flat file:**

- Avoids data duplication.
- Avoids inconsistent records
- Easier to change data
- Easier to change data format
- Data can be added and removed easily
- Easier to maintain security.

## 2. Problem Statement

To be able to index cases in the existing system, the first approach was to process the spreadsheets provided by CNLG which contain information about persons' trials and the court names.

During that process it appeared that around 50% of data provided contain errors based on a mismatch between the physical court names and the one provided in the spreadsheet. This situation had for consequence going into a physical check to ensure data integrity and quality.

Out of this, it also appeared that most of the data which could have been imported previously had many duplicates[1] based on qualified parameters.

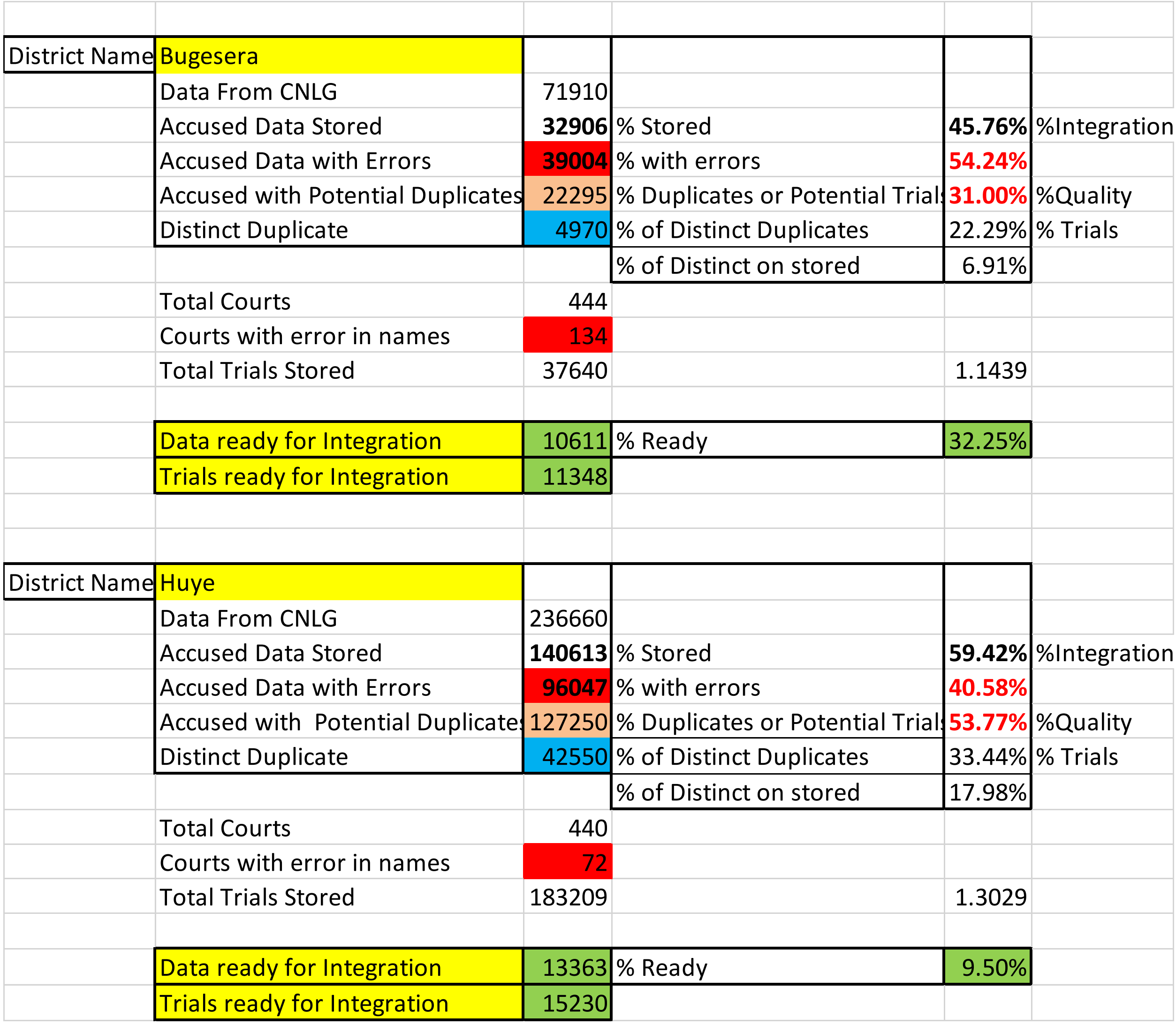
## 3. Approach

1. To correct these errors and come up with accurate information to be handled by the system, a customised tool has been created to identify, correct errors and potentially identify duplicate records based on qualified parameters.

[1] Duplicate : An accused person having matching names that appear more than one time.

Based on those records, **only** non-duplicates data were successfully integrated into the indexing platform. All the others are still pending for a defined strategy based on our recommendation.

## 4. Findings



| | | | | | |
|---|---|---|---|---|---|
| District Name | Bugesera | | | | |
| | Data From CNLG | 71910 | | | |
| | Accused Data Stored | **32906** | % Stored | **45.76%** | %Integration |
| | Accused Data with Errors | **39004** | % with errors | **54.24%** | |
| | Accused with Potential Duplicates | 22295 | % Duplicates or Potential Trial | **31.00%** | %Quality |
| | Distinct Duplicate | 4970 | % of Distinct Duplicates | 22.29% | % Trials |
| | | | % of Distinct on stored | 6.91% | |
| | Total Courts | 444 | | | |
| | Courts with error in names | 134 | | | |
| | Total Trials Stored | 37640 | | 1.1439 | |
| | | | | | |
| | Data ready for Integration | 10611 | % Ready | 32.25% | |
| | Trials ready for Integration | 11348 | | | |
| | | | | | |
| | | | | | |
| District Name | Huye | | | | |
| | Data From CNLG | 236660 | | | |
| | Accused Data Stored | **140613** | % Stored | **59.42%** | %Integration |
| | Accused Data with Errors | **96047** | % with errors | **40.58%** | |
| | Accused with Potential Duplicate | 127250 | % Duplicates or Potential Trial | **53.77%** | %Quality |
| | Distinct Duplicate | 42550 | % of Distinct Duplicates | 33.44% | % Trials |
| | | | % of Distinct on stored | 17.98% | |
| | Total Courts | 440 | | | |
| | Courts with error in names | 72 | | | |
| | Total Trials Stored | 183209 | | 1.3029 | |
| | | | | | |
| | Data ready for Integration | 13363 | % Ready | 9.50% | |
| | Trials ready for Integration | 15230 | | | |

Image above shows that two districts have been processed (Bugesera and Huye). The two districts have 71,910 and 236,660 trials respectively.

During the process of identifying errors, among the stored data, we observed a big percentage of duplicates, the reason why there is a low number of data ready for integration.

The following was observed:

1. High percentage of duplicates data. Average 50% of the entire data set.
2. We realized that the information contained in spreadsheet provided by CNLG is more about trial cases than it is for accused persons and of a particular case. Since, originally these data were recorded from different reports of Courts' activities.

This is justified by the fact that based on the number of trial cases, an accused appears more than once, and could have been tried in different courts (cell, sector or appeal).

**Example:** BAKUNDUKIZE Vincent

**a. Cell court level:**

| UMURENGE | IN_RUS_UMURENGE | GACACA_AKAGARI | UWABURANYE | IGITSINA | UBWENEGIHUGU | UMURIMO 1994 | SE | NYINA | UMWAKA_AMAVUKO | ITARIKI_URUBANZA | IGIHANO | URWEGO YABURANIYEMO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KIGOMA | BITARE | FUKWE | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA | NYINAWABANYIGINYA | - | 31/10/2009 | 4 300 FRW | URWEGO RWA 3 |
| KIGOMA | BITARE | FUKWE | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA | NYINAWABANYIGINYA | 1954 | 15/8/2007 | 4 300 FRW | URWEGO RWA 3 |
| KIGOMA | BITARE | FUKWE | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA | NYINAWABANYIGINYA | - | 31/10/2009 | 4 300 FRW | URWEGO RWA 3 |
| KIGOMA | BITARE | FUKWE | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA | NYINAWABANYIGINYA | 1954 | 15/8/2007 | 4 300 FRW | URWEGO RWA 3 |

**b. Sector court level**

| UMURENGE | IN_RUS_UMURENGE | GACACA_AKAGARI | UWABURANYE | IGITSINA | UBWENEGIHUGU | UMURIMO 1994 | SE | NYINA | UMWAKA_AMAVUKO | ITARIKI_URUBANZA | IGIHANO | URWEGO YABURANIYEMO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KIGOMA | BITARE B | | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA | MUJAWABANYIGINYA | 1954 | 28/11/2007 | UMWERE | URWEGO RWA 2 |

**c. Appeal court level**

| UMURENGE | IN_RUS_UMURENGE | GACACA_AKAGARI | UWABURANYE | IGITSINA | UBWENEGIHUGU | UMURIMO 1994 | SE | NYINA | UMWAKA_AMAVUKO | ITARIKI_URUBANZA | IGIHANO | URWEGO YABURANIYEMO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KIGOMA | BITARE B | | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA FRANCOIS | | | 27/4/2007 | 7 400 FRW | URWEGO RWA 3 |
| KIGOMA | BITARE B | | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA FRANCOIS | | | 18/2/2008 | 2 968 FRW | URWEGO RWA 3 |
| KIGOMA | BITARE B | | BAKUNDUKIZE VINCENT | GABO | RWANDA | UMUHINZI | NZARAMBA FRANCOIS | | | 27/8/2008 | 16475 FRW | URWEGO RWA 3 |

Given the example above, a person called Bakundukize Vincent appears:

*4 times in the Cell court list,

*Once in a Sector court list

* And 3 times in an appeal court list. In total, Mr Bakundukize appears 8 times on the list.

For the digitization and indexing process this can’t be processed as we will end with a digital repository with poor data quality – and loose its integrity.

## 5. Recommendations

- In the above case and other similar cases, the key task is to clean and harmonize and merge all data to make an accused with a unique identifier. The accused should appear only once to be later on linked to multiple cases he has been involved in.
- At the end of this process, we will have an exhaustive list of all accused persons per court and this list will be extracted from the existing case trial spreadsheet.
- Having a big amount of duplicates data, we should definitely come up with a solution that merges all those records into distinct one based on pre-defined qualified parameters inside a tool (merger service).
- Once an accused list is created and duplicates resolved, it will be much easier and quicker to assign cases to an accused. If the accused is not available on the list, the system will give the possibility to add or create a new accused.
- This means that the exact number of accused is likely to reduce once the duplicates are handled in the above explained way.
- The exact figures of trials will be known at the end of the indexing process once all cases have been assigned to their accused persons.
- To measure success of indexing based on a number of steps such as number of accused with a unique identifier , number cases linked to any accused , data fully integrated with errors and number accused linked to trial process