Executive Summary

This document aims to act as a log of all technical activities and thought processes on the implementation of the AFIS solution for the INEC state database and the national voters database.

Understanding How AFIS Works

Quick Definitions

AFIS	Automated Fingerprint Identification System
ClusterSix	Suite of solutions, including AFIS installed on a single server or on clustered servers
100 bytes	The size of a single comparison between two fingers
1k	One Kilobyte, the average size of the AFIS results between any two individuals with 10 fingers each, comparing on fingers in the same positions e.g Right Thumb to Right Thumb, RI to RI, RM to RM, RR to RR, RL to RL and storing the results of the same
Subject	A record in the database which is an individual with fingers. This may be simply referred to as *individual* or *record* in this document.

Similarity Score	A final figure calculated based on the amount of common features shared between the fingerprints of two individuals. The higher this figure, the more higher the probability that these individuals are duplicates.
Comparisons	The number of individual to individual similarity scores
Calculations	The number of comparisons multiplied by 10 (10 fingers) from which the similarity score is calculated.

AFIS Modes

The AFIS algorithm identifies potential duplicates by searching for similarities between each individual and other individuals in the database.

AFIS can be executed in several modes and this is commonly done in multiple phases due to the fact that it is a compute intensive task and long running process.

Blanket AFIS

A **blanket** AFIS algorithm calculates the similarities between every individual and every other individual in the database.
Under this mode, value for the number of comparisons is calculated as follows

```
Number of Comparisons C = (number of individuals) * (number of individuals)
```

This can be rewritten as

```
C = n * n
```

Since it makes no sense to compare a record against the exact same record, this formula is updated to

```
C = n * (n - 1)
```

Furthermore, comparing individual A with individual B is exactly the same as comparing individual B to A, which means we will have half the total instead.

C = n * (n - 1) * ½

or **C = 0.5n * (n - 1)**

where n is the number of individuals in the database.

Actual figures:

If we have 10 individuals in the database, the number of comparisons, C will be

$$C_{10} = 10 * (10 - 1) * ½$$

$$C_{10} = 45$$

If we have 100 individuals

$$C_{100} = 100 * (100 - 1) * ½$$

$$C_{100} = 4,950$$

If we have 1000 individuals

$$C_{1000} = 1000 * (1000 - 1) * ½$$

$$C_{1000} = 499,500$$

If we have 10000 individuals

$$C_{10000} = 10000 * (10000 - 1) * ½$$

$$C_{10000} = 49,995,000$$

BioProfile-Based AFIS

A **bioprofile-based** AFIS algorithm calculates the fingerprint similarities between every individual and every other individual who share commonalities in a number of non-fingerprint related characteristics. The most useful of these are Age and Gender. This compares individuals only with the same gender and within the same age group. Contrasting this with the Blanket AFIS mode, this mode results in a drastically lower number of comparisons.

Gender Filter

Using gender alone, the number of comparisons becomes
instead of

$$C = n * (n - 1) * ½$$

With Gender -

$$C = (n_m * (n_m - 1) * ½) + (n_f * (n_f - 1) * ½)$$

We have a 53% male, 47% female gender distribution across the 73 million people-database. However, for the purposes of simplification we will assume a 50% male - 50% gender distribution.

Therefore this now becomes

$$C = (n/2 * (n/2 - 1) * ½) + (n/2 * (n/2 - 1) * ½)$$

$$C = (n/4 * (n/2 - 1)) + (n/4 * (n/4 - 1))$$

$$C = n/4 * ((n/2 - 1) + (n/2 - 1))$$

$$\mathbf{C = n/4 * (n - 2)}$$

Individuals	Blanket AFIS	BioProfile (Gender)	Difference	% Reduction
10	45	20	25	55.6%
100	4950	2450	2,500	50.5%
1,000	499,500	249,500	250,000	50.1%
10,000	49,995,000	24,995,000	25,000,000	50.0%
100,000	4,999,950,000	2,499,950,000	2,500,000,000	50.0%

Using a Gender Filter reduces the number of comparisons by half.

Age Group Filter

Here, we assume that the likelihood that a registration will be a duplicate of another registration outside his/her age group is negligible.
We predetermine that age group should be calculated against the individual's registered date of birth and the 'radius' of this age group is configurable.

If 10 is the radius of the age group, an individual who is 35 years old will only be compared against other individuals in the database who are between the ages of 25 and 45 years.

Similarly, if 15 is the radius, a 35 year old man will only be compared with individuals who are between 20 and 50 years old.

By specifying that the individuals in my target group should be within my age radius, the number of calculations will be determined as follows.

$c^1 = n^1 * (n^1 - 1) * ½$
Where
c^1 is the number of comparisons with the subject (individual a)
n^1 is the number of individuals within the age radius of individual a

The total number of comparisons is across the database is therefore

$C' = [(n^1 * (n^1 - 1) * ½) + (n^2 * (n^2 - 1) * ½) + ... + (n^m * (n^m - 1) * ½)]$

where **m** is the last in the list of individuals.
While this may look complicated, this is a fairly straightforward and results in a drastically smaller set of comparisons.

Proximity Filter

Better known as **Electoral Advantage,** this questions the possibility that an individual would be willing or privileged to travel very long distances to register multiple times and also be able to cast his/her vote on election day in the two different and positionally distant locations. This is especially aided by the fact that as there is in effect a **no-movement order** on the **election day**. Therefore the further we move (in kilometres) from a voting station, the less the likelihood of duplication.

Using the GPS coordinates of the voting stations, we are able to calculate, from a pivot polling unit, the distance between that polling unit and every other polling unit in the database. If we set the radius to 20 kilometres, it means that the duplicate search parameters will only focus on individuals who registered within 20 kilometres in any direction of my own polling unit.

Using Bayelsa State as a case study for these calculations, this is the summary of the findings using the Proximity Filter

Individuals	Blanket Comparisons	In 50Km Radius	Difference	%*
640,459	205,093,545,111	59,415,831,197	145,677,713,914	71%

Individuals	Blanket Comparisons	In 10Km Radius	Difference	%*
640,459	205,093,545,111	5,426,908,768	199,666,636,343	97.4%

*% reduction in the number of comparisons.

Hybrid Filter

This is a combination of the proximity filter, gender filter and age group filter together to form the basis of comparisons. This focuses on the most likely group of individuals who may be the duplicate of an individual - someone who shares my gender, is within my age group and also registered within 10 kilometres of where I registered.

This in effect will essentially create a subset of calculations which will give an efficient way of executing the deduplication process.

By extrapolating the figures above, this means that

This needs to be verified with tests!

Individuals	Blanket Comparisons	Hybrid Filtered (10KM)	Difference	%*
640,459	205,093,545,111	814,036,316	204,279,508,795	99.6%