Pathway to realizing the Aadhaar promise: Government of Maharashtra’s UID Innovation Centre

Developed For:
Directorate of Information Technology,
Government of Maharashtra
Author: Murtuza Karachiwala, Ernst & Young Private Limited

Table of Contents

1. Abstract

In the words of the Chairman of the UIDAI, the name Aadhaar communicates the fundamental role of the unique number issued by the UIDAI as a universal identity infrastructure, a foundation over which public and private agencies can build services and applications that benefit residents across India. In this case study, we have attempted to document the learning of Government of Maharashtra and to discuss the fundamental challenges, pre-requisites and initiatives required for realizing the true transformational potential of Aadhaar.

Maharashtra State has been amongst the pioneers in creation of the State Resident Data Hub (SRDH). However, the implementation of the Direct Benefit Transfer Scheme (DBTS) and other attempts for Aadhaar integration unveiled several challenges that required innovative solutions. In a unique initiative, the Directorate of Information Technology has set-up the UID Innovation Center for addressing these challenges. Under the leadership of Shri. Rajesh Aggarwal, Secretary (IT), the UID Innovation Center is currently undertaking unparalleled data quality, cleansing and integration initiatives to make the contents of SRDH usable.

Some of the components, key challenges and learning discussed in the case study include:

1. Infrastructure and resources required for handling large volumes of residents’ data
2. Need for SRDH data standardization – especially of address data
3. Data mining and analysis for development of customized data cleaning, normalization and simplification rules
4. Data de-duplication on demographic parameters in eGovernance systems and beneficiary lists
5. Data matching and seeding for Aadhaar integration
6. Use of KYR details for identification of potential ‘ghost entries’ in beneficiary lists
7. Verification of authenticity of bank account details for Direct Benefit Transfer
8. Maintaining AUA ASA history to reduce dependency on the Central ID repository (CIDR)

One of the key mandates and role of the UID Innovative center is to create re-usable components and solutions which can be leveraged for overcoming similar challenges that would be faced by other states and private agencies. The case study provides details with specific examples for each of the challenges. The following re-usable components and solutions developed by the UID Innovation would be made available on specific request made to the Directorate of Information Technology, Government of Maharashtra:

1. Algorithm / code for verification of UID numbers in eGovernance systems and beneficiary lists
2. Algorithm / code for data mining
3. Outputs from data mining – including character and token and frequency analysis
4. Algorithm / code for cleaning of English Names
5. Algorithm / code for cleaning of Marathi Names
6. Algorithm / code for normalization and simplification of English names for matching / seeding

2. Key words and tags

Sr. No.	Key words and tags	Description
1	Aadhaar	Refers to the Unique Identity or Aadhaar number
2	ASA	Refers to Authentication Service Agency
3	AUA	Refers to Authentication User Agency
4	CIDR	Refers to the Central Identity Repository set-up by UIDAI in Bangalore
5	DBTS	Refers to the Direct Benefit Transfer Scheme
6	DIT	Refers to the Directorate of Information Technology, Government of Maharashtra
7	KYR	Demographic details (identity related information) of the resident
8	KYR+	Additional data about the resident that pertains to non-identity information which has been mapped to his Aadhaar
9	NDSAP	Refers to the National Data Sharing and Accessibility Policy
10	SRDH	Refers to the State Resident Data Hub
11	UID	Refers to the Unique Identity or Aadhaar number
12	UIDAI	Refers to the Unique Identity Authority of India
13	U-SRDH	Refers to the Usable State Resident Data Hub

3. Note to Practitioners/Instructors

The case study is meant for State Authorities, IT vendors doing eGovernance work related to Aadhaar integration and other private agencies to understand the key lessons learnt and innovations required for overcoming some of the basic challenges to initiatives required for leveraging Aadhaar as the common identity infrastructure across systems, applications and organizations that interact with residents of India.

The following re-usable components and solutions developed by the UID Innovation would be made available on specific request made to the Directorate of Information Technology, Government of Maharashtra:

1. Algorithm / code for verification of UID numbers in eGovernance systems and beneficiary lists
2. Algorithm / code for data mining
3. Outputs from data mining – including character and token and frequency analysis
4. Algorithm / code for cleaning of English Names
5. Algorithm / code for cleaning of Marathi Names
6. Algorithm / code for normalization and simplification of English names for matching / seeding

4. Project Context

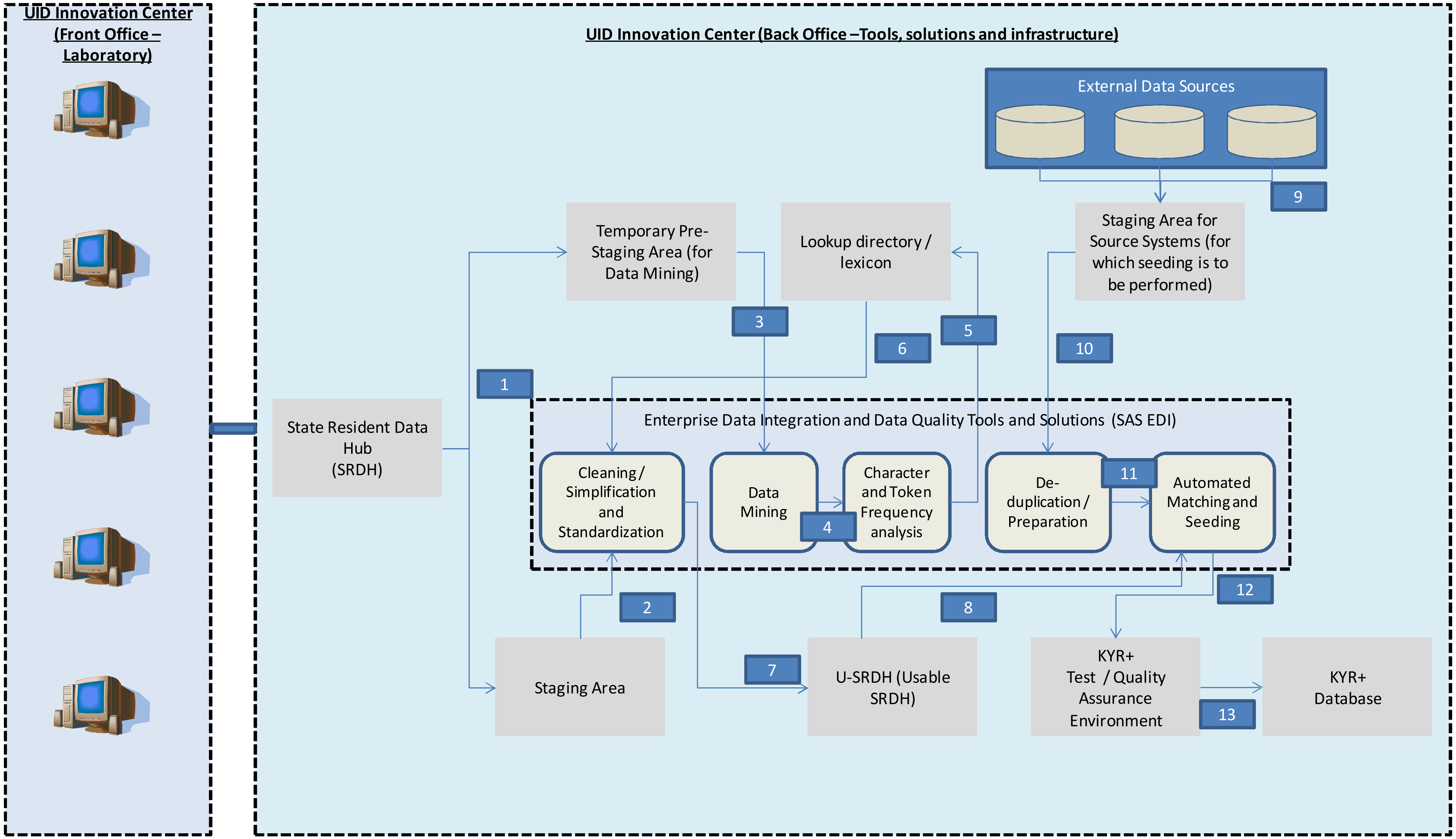
'Aadhaar' communicates the fundamental role of Unique Identity Number as a universally accepted identifier for interacting and transacting with a resident of India. Traditionally, the Name of an individual has been among a set of attributes used to perform this function. Hence, it is critically important to have a context-rich understanding of patterns in data – especially the name – from eGovernance systems to be able to transition the traditional methods to Aadhaar based systems.

The UID Innovation Center, set-up by Government of Maharashtra, has been at the fore-front for development of innovative solutions to enable a smooth transition. The activities of the UID Innovation Center consist of a program of several interlinked research projects and experiments being managed to achieve this common objective. It is undertaking several initiatives for data quality, cleansing and integration to make Aadhaar the basis for identity across systems.

5. Project Overview

Maharashtra State has been amongst the pioneers in creation of the State Resident Data Hub (SRDH). However, the implementation of the Direct Benefit Transfer Scheme (DBTS) and other attempts for Aadhaar integration of key databases managed by the State Government unveiled several challenges that required innovative solutions. In a unique initiative, the Directorate of Information Technology has set-up the UID Innovation Center for addressing these challenges.

The overview of the key components and processes that constitute the UID Innovation center are depicted in the figure shown below:


UID Innovation Center (Front Office – Laboratory)
UID Innovation Center (Back Office –Tools, solutions and infrastructure)
External Data Sources
9
Staging Area for Source Systems (for which seeding is to be performed)
Temporary Pre-Staging Area (for Data Mining)
Lookup directory / lexicon
3
5
6
10
1
State Resident Data Hub (SRDH)
Enterprise Data Integration and Data Quality Tools and Solutions (SAS EDI)
Cleaning / Simplification and Standardization
Data Mining
4
Character and Token Frequency analysis
De-duplication / Preparation
11
Automated Matching and Seeding
12
2
8
Staging Area
7
U-SRDH (Usable SRDH)
KYR+ Test / Quality Assurance Environment
13
KYR+ Database

■ -- Data Processes (See Explanation below)

1. Extraction, Transformation and Loading of raw data
2. Cleaning and Simplification of Names
3. Data Mining and Token Generation
4. Character and Token frequency analysis
5. Token combination analysis for local language
6. Enrichment of data in localized language
7. Data Standardization and enrichment
8. Match Code generation and de-duplication
9. Source data loading
10. Source data pre-processing, de-duplication and standardization
11. Match code generation and development of matching rules
12. Manual Quality Assurance and data testing
13. Exposing KYR+ data

As depicted above the major data processes that are currently underway at the UID Innovation Center include the following:

1. Extraction, Transformation and Loading of raw data: Having set-up the State Resident Data Hub (SRDH), the Directorate of Information Technology, Government of Maharashtra realized that there were several initiatives that would be required to make the data useful. The SRDH data was not standardized and had several data quality issues. Some of the key issues included the following:
   a. The address fields in the SRDH data were not standardized.
   b. There were several errors in come of the data fields.
   c. Marathi Language fields were not standardized and transliteration was not effective

   Hence, it was required that this "raw" data should be transformed to make a clean, validated copy that could be used for further data processes. The UID Innovation center has implemented a scheduled process for this Extraction, Transformation and Loading of raw data. This process takes the data in the SRDH and loads a clean, validated copy into a staging area.

   The UID Innovation center also conducted experiments to conclude that a set of customized rules would have to be developed through the discovery of patters in the existing raw data. For this mining and analysis a copy of the raw data was made in a temporary Pre-Staging Area. Hence, this process helps isolate the SRDH from activities of the UID Innovation Center and helps maintain integrity and security of the SRDH.

2. Cleaning and Simplification of Names: In some of the experiments conducted by the UID Innovation center, the team realized that there were limited quality controls for data entry during UID enrolment. There were several errors and inconsistencies - particularly in the names of individuals. Hence, a set of rules were developed to clean and simplify the names in the SRDH. These set rules keep evolving through the outputs of processes 3, 4, 5 and 6 below.

3. Data Mining and Token Generation: This is the first in a set of processes that aim to discover patterns and develop cleaning and simplification of names in the SRDH database (see process 2 above). In its first iteration the UID Innovation center was working with names of 5.2 Crore residents in English and Marathi. To be able to meaningfully analyze and detect patterns from this large dataset, it was required to perform mining of this data and generate tokens. Through this process, only 12 lakh distinct tokens were found in English names and 13 distinct tokens in Marathi. This depicts a major compression in the data for processing further.

4. Character and Token frequency analysis: During generation of the tokens, the frequency of occurrence of each token was recorded. The top ten tokens in English and Marathi are as shown below:

Table 1: Top 10 Most Frequent Occurring Tokens (In English and Marathi Names)

Token in English	Frequency	Token in Marathi	Frequency
shaikh	1350831	शेखी	1623166
patil	846554	पाटीली	844975
khan	623086	खानी	613773
jadhav	543412	जाधवी	558362
sanjay	458220	मोहम्मदी	521028
ashok	454412	संजयी	462369
pawar	421571	अशोकी	453588
suresh	412470	पवारी	430646
ramesh	403607	सुरेशी	408698
shinde	398985	रमेशी	404434

The tokens were then further analyzed to generate output that depicted the frequency of each character, each combination of 2 characters and each combination of three characters – in the beginning, middle and end of a token. The top 5 combinations of characters in English and Marathi are as shown below:

Table 2: Top 5 most frequently occurring characters in Names (English)

Character	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
a	7081124	116878716	12243946	136203786
S	17071920	34106061	1471772	52649753
h	1733466	38233074	8196800	48163340
r	6527993	32487554	8239039	47254586
n	3634325	31994589	6740898	42369812

Table 3: Top 5 most frequently occurring characters in Names (Marathi)

Character	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
◌ी	0	50101654	10227095	60328749
री	6449891	30422765	9737383	46610039
मी	7407346	18258526	4073293	29739165
सी	10360160	15613211	1483337	27456708
वी	4498481	14536840	4825999	23861320

Table 4: Top 5 most frequently occurring 2-character combinations in Names (English)

2 Character combination	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
an	1614016	16321433	4915606	22851055

2 Character combination	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
ha	1077662	18433126	1500814	21011602
Sh	6688065	7359110	4385734	18432909
Ra	5326463	11277439	1819763	18423665
ar	681412	8931219	6872444	16485075

Table 5: Top 5 most frequently occurring 2-character combinations in Names (Marathi)

2 Character combination	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
री	3475113	6539211	390604	10404928
्री	0	5011755	1448686	6460441
ारी	0	3982935	2067413	6050348
ंदी	0	3808004	770170	4578174
नी	3475113	6539211	390604	10404928

Table 6: Top 5 most frequently occurring 3-character combinations in Names (English)

3 Character combination	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
Sha	3619324	1413735	615929	5648988
han	191002	3503692	1612068	5306762
Ram	1778762	358865	1992404	4130031
kar	252575	739519	3044992	4037086
and	12439	3319627	469117	3801183

Table 7: Top 5 most frequently occurring 3-character combinations in Names (Marathi)

3 Character combination	Occurrences at the Start Of Name	Occurrences in the Middle of Name	Occurrences at the End Of Name	Total Occurrences
राम	1238102	106186	1925470	3269758
राव	102723	128444	2657276	2888443
	2153986	193991	0	2347977
	15672	628210	1442451	2086333
ं ६	0	1953031	0	1953031

Such output (developed for the entire SRDH data) was then manually analyzed to detect patterns, exceptions and invalid combinations which need to be corrected. A customized set of cleaning and simplification rules where developed for both English and Marathi Names.

5. Token combination analysis for local language: It was also observed that local language (Marathi) data was not in sync with the English data. The tokens were correlated to create a

look-up directory of English-marathi combinations. Some examples of token name correlation are as shown below:

Table 8: Example of correlation in English-Marathi Names

English Name	Marathi Name	Frequency
Akshay	अक्षय	68337
Akshay	आक्षय	37
Akshay	Akshay	32
Akshay	अक्षशय	4
Akshay	`अक्षय	2
Akshay	अक्षयकुमार	2
Akshay	अक्ष शय	1
Akshay	Akshy	1
Akshay	अक्षयसिंह	1
Akshay	अ`क्षय	1
Akshay	अकक्षय	1
Akshay	अ क्षय	1
Akshay	अक्षया	1
Akshay	अक्षी	1
Akshay	आ`क्षय	1
Akshay	आशय	1
Akshay	एक्षय	1
Akshay	य	1
Akshay	वलास	1

This output was developed across all tokens in the SRDG database. In the example above, ideally there should be only one way in which the name 'Akshay' can be written in Marathi. This output was used as a look-up directory for cleaning and simplification of Marathi names.

6. Enrichment of data in localized language: The directory created in process 5 above was then used to enrich the Marathi data by replacing the less frequent combinations with the most frequent value for a given English token.

7. Data Standardization and enrichment: Using advanced tools and algorithms, the data in the staging areas was then standardized, enriched and loaded into the Usable State Resident Data Hub (U-SRDH). The data in the U-SRDH would be the base for any further processes and uses – including for automated seeding and web services.

8. Match Code generation and de-duplication: The cleaned names in the U-SRDH database are then used to generate "match codes" which allow for fuzzy matching to names for automated seeding. Once the match codes are generated, the logic could also be used for demographic de-

duplication of the UID database and for seeding / matching with other databases. Examples of duplicates include the following:

Table 9: Examples of potential duplicate records in SRDH

UID	Name	Date of Birth	Age	Gender	Village Name	District Name
695XXX264	Ganapat Shankar Patil	01-01-1932	81	M	Narangi	Raigarh
970XXX222	Ganapat Shankar Patil	01-01-1932	81	M	Narangi	Raigarh
776XXX045	Shabana Bagwan	01-01-1981	32	F	Jalna	Jalna
919XXX786	Shabana Bagwan	01-01-1981	32	F	Jalna	Jalna
775XXX106	Rafik Sayyad	01-01-1991	22	M	Jalna	Jalna
418XXX475	Rafik Sayyad	01-01-1991	22	M	Jalna	Jalna
696XXX009	Sultana Shaikh	01-01-1971	42	F	Jalna	Jalna
441XXX798	Sultana Shaikh	01-01-1971	42	F	Jalna	Jalna
696XXX777	Gulab Haribhau Dahole	03-01-1966	47	M	Shivaji Nagar S.O	Amravati
866XXX863	Gulab Haribhau Dahole	03-01-1966	47	M	Shivaji Nagar S.O	Amravati
714XXX760	Sultana Khan	01-01-1981	32	F	Jalna	Jalna
865XXX139	Sultana Khan	01-01-1981	32	F	Jalna	Jalna
716XXX069	Rahul Jadhav	01-01-1989	24	M	GHANSWANGHI	Jalna
575XXX910	Rahul Jadhav	01-01-1989	24	M	GHANSWANGHI	Jalna
750XXX393	Bhau Rathod	01-01-1994	19	M	PARTUR	Jalna
261XXX794	Bhau Rathod	01-01-1994	19	M	PARTUR	Jalna
773XXX656	Parvatibai Kete	01-01-1961	52	F	naigaon	Nanded
667XXX356	Parvatibai Kete	01-01-1961	52	F	naigaon	Nanded
748XXX239	Shifa Shaikh	01-01-2009	4	F	govandi	Mumbai

UID	Name	Date of Birth	Age	Gender	Village Name	District Name
218XXX333	Shifa Shaikh	01-01-2009	4	F	govandi	Mumbai
716XXX100	Dilshad Shaikh	01-01-1966	47	F	Solapur North	Solapur
248XXX837	Dilshad Shaikh	01-01-1966	47	F	Solapur North	Solapur
724XXX024	Dilshad Shaikh	01-01-1966	47	F	Solapur North	Solapur
804XXX943	Dilshad Shaikh	01-01-1966	47	F	Solapur North	Solapur

A total of 9270 records were found, for which potential duplicates existed in the system.

9. Source data loading: To ensure security and integrity of the source data, a separate secure staging area - where data from other external systems can be validated and processed for comparison and automated seeding with SRDH data - has been created.

10. Source data pre-processing, de-duplication and standardization: Once the source data has been loaded, a set of pre-processing de-duplication and standardization rules are run on the source data sets. The objective of this process is to be able to identify key challenges in the source data so that suitable workarounds can be implemented before seeding is done. The process also enables DIT to provide feedback to the concerned department for improvement in data quality of the source systems. For example, during pre-processing of the election database the following issues were highlighted:

    - There are 6097 records in which the age of the citizen is less than 18 years.
    - There are 3,26,087 records where age of the citizen is greater than 90 years. 217 of these records are more than 150 years old, with the maximum age recorded as 990 years.
    - There are approximately 1.3 Crore records which have card ID as null or blank.
    - There were 11.81 lakh duplicate records found in the database. Some examples of duplicates are as shown below:

Table 10: Example of duplicates in election database

Assembly No	House No	Voter Name	Relative s First Name	Relative s Last Name	Sex	Age
287	8	Sushila Pandurang Patil	Pandurang	Patil	F	59
287	8	Girish Pandurang Patil	Pandurang	Patil	M	37
287	100	Sushila Pandurang Patil	Pandurang	Patil	F	59
287	100	Girish Pandurang Patil	Pandurang	Patil	M	37
287	150	Kiran Ramchandra Patil	Ramchandra	Patil	M	38
287	150	Kishor Ramchandra Patil	Ramchandra	Patil	M	37
287	150	Malutai Ramchandra Patil	Ramchandra	Patil	F	62

Assembly No	House No	Voter Name	Relative s First Name	Relative s Last Name	Sex	Age
287	180	Kiran Ramchandra Patil	Ramchandra	Patil	M	38
287	180	Kishor Ramchandra Patil	Ramchandra	Patil	M	37
287	180	Malutai Ramchandra Patil	Ramchandra	Patil	F	62

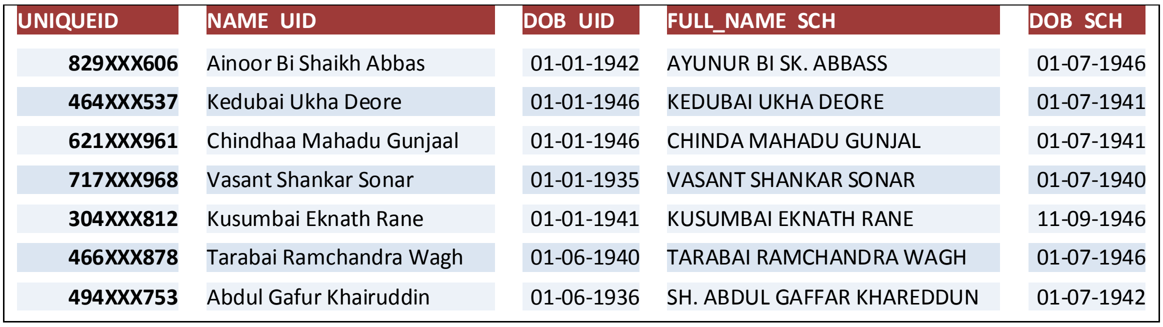
11. Match code generation and development of matching rules: Once the source data is ready for seeding, the match code generation algorithms is tuned and run. At times multiple runs may be required for maximize the percentage of records that result in high probability matches. The algorithms are then run on the entire dataset to perform automated seeding. Examples of the results of automated seeding and statistics for some of the initial seeding jobs are as shown below:

Table 11: Statistics on first few seeding jobs

District	Scheme name	No of records in Input file	Name and DOB matches	Name Matches	Percentage of High Probability matches	Percentage of Matches	Enrolment % (% of population enrolled)
Mumbai	SSJD_Scholarship -OBC	8861	2810	4830	31.71%	54.51%	58.00%
Mumbai	SSJD_Scholarship -SC	10993	4059	6903	36.92%	62.79%	58.00%
Nandurbar	Scholarship -OBC	8334	1217	3780	14.60%	45.36%	21.00%
Nandurbar	Scholarship -SC	2603	387	963	14.87%	37.00%	21.00%
Wardha	Scholarship -OBC	27248	9678	18865	35.52%	69.23%	84.00%
Wardha	Scholarship -SC	14654	4646	9155	31.70%	62.47%	84.00%
Pune	Scholarship -OBC	30131	3943	10926	13.09%	36.26%	33.00%

Table 12: Illustrative automated seeding output

UNIQUEID	NAME_UID	DOB_UID	FULL_NAME_SCH	DOB_SCH
293XXX957	Bhamabai Shankar Aher	01-06-1940	BHIMRAO SHANKAR AHIRE	01-07-1940
973XXX778	Sarubai Namdev Ahire	01-01-1946	SARUBAI NAMDEO AHIRE	01-07-1946
753XXX059	Shaikh Abbas Shaikh Ismail	01-01-1937	SHAIKH ABBAS SHAIKH USMAN	01-07-1936
582XXX884	Zubaida Gayasuddin	01-01-1944	JUBEDA GAYASUDDIN	01-07-1943
351XXX628	Yamunabai Dattatrey Jadhav	01-01-1946	YAMUNABAI DATTATRAY JADHAV	13-09-1944
488XXX674	Subhabai Popat Borse	01-01-1946	SUBHABAI POPAT BORSE	01-07-1944
359XXX342	Banyabai Shamrao Sonawane	01-01-1946	BANYABAI SHAMRAO SONAWANE	17-11-1943
223XXX938	Indubai Anandaa Aahire	01-01-1943	INDUBAI ANANDA AHIRE	01-07-1945
896XXX040	Vasant Dhana Mali	01-03-1943	VASANTRAO DHANA MALI	01-07-1940



UNIQUEID	NAME_UID	DOB_UID	FULL_NAME_SCH	DOB_SCH
829XXX606	Ainoor Bi Shaikh Abbas	01-01-1942	AYUNUR BI SK. ABBASS	01-07-1946
464XXX537	Kedubai Ukha Deore	01-01-1946	KEDUBAI UKHA DEORE	01-07-1941
621XXX961	Chindhaa Mahadu Gunjaal	01-01-1946	CHINDA MAHADU GUNJAL	01-07-1941
717XXX968	Vasant Shankar Sonar	01-01-1935	VASANT SHANKAR SONAR	01-07-1940
304XXX812	Kusumbai Eknath Rane	01-01-1941	KUSUMBAI EKNATH RANE	11-09-1946
466XXX878	Tarabai Ramchandra Wagh	01-06-1940	TARABAI RAMCHANDRA WAGH	01-07-1946
494XXX753	Abdul Gafur Khairuddin	01-06-1936	SH. ABDUL GAFFAR KHAREDDUN	01-07-1942

12. Manual Quality Assurance and data testing: Once the results of the automated seeding are generated, the records are scrutinized manually to identify issues. If there are opportunities to further tune the matching algorithms, then process 11 is repeated.

13. Exposing KYR+ data: The final seeded output is put in the Know-your-resident (KYR+) database. Several web services are currently being developed in the KYR+ database to enable access through web services. For example, AUA ASA authentication history is maintained as part of the KYR+ database. Departments that do not need a real time authentication (where it is adequate that identity of a resident has been authenticated within a particular time frame – say 1 month) can use the authentication history instead of performing CIDR based authentication for every transaction.

6. Issues and challenges

Some of the key issues and challenges that have been encountered and overcome by the UID Innovation Center include the following:

- Infrastructure and resources required for handling large volumes of residents' data: Traditional eGovernance systems are not built to handle large volumes of data and complex processing such as character and token frequency analysis and automated seeding. Hence, atypical data handling algorithms and distributed resources are necessary for the data processes being executed by the back-office of UID Innovation Center. For Maharashtra, the use of cloud computing (implemented through the Maharashtra State Data Center) enable dynamic allocation of resources as and when required. Multiple virtualized servers with separate database instances were used to make sure that parallel data processing and aggregation of results is possible.

- Data mining and analysis for development of customized data cleaning, normalization and simplification rules: Aadhaar integration through seeding of UID is a multi-step process. The first of these steps is to first match the names and other demographic data in both the databases (UID database as well as Beneficiary Database). There may be several software based approaches that can be used to automate the process of matching and seeding. However, the use of generic rules may limit the extent to which these rules can be successful. For example, in Wardha (district in Maharashtra State) basic, generic software algorithms were used for getting results between 30% to 50% of the records being matched. During the pilot runs and experiments for implementation of automated seeding conducted by the UID innovation center, it was evident that generalized rules for data cleaning and matching may not be able to provide matching output for more than 50% of the records. Hence, there was a need to develop customized rules based on discovery of patterns in the existing data. The names of individuals in Maharashtra had a few peculiar characteristics that were specific to the local context and culture. For example, the use of the suffix "rao" or "tai" at the end of names. To develop these rules, specific data mining activities were undertaken and look-up directories were frequency distributions of the names and character combinations were created.

Also, particular for the data in local language, i.e. Marathi, specific rules had to be developed to enable cleansing of data – especially names. These customized rules (through algorithms and code) will be made available through the Public data portal on the Aadhaar Maharashtra website.

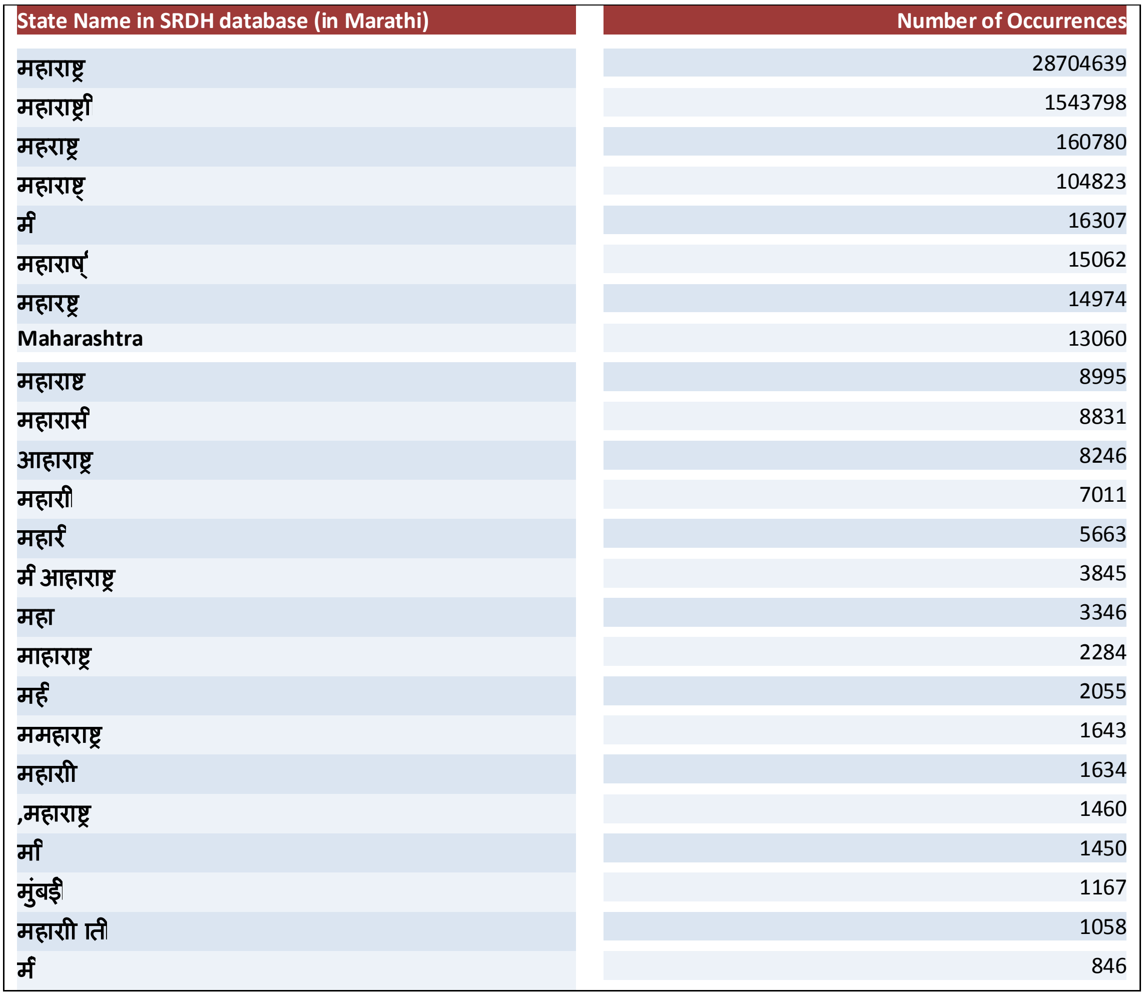
7. Key Lessons

Some of the key lessons learned from the experiences at the UID Innovation Center include the following:

- Need for SRDH data standardization – especially of address data: The Government machinery in most states is organized as a hierarchy which has well-defined authorities for roll-out of mandates at the ground level. At a village level, specific officials can be made in-charge of a particular mandate. Hence, any control measures and initiative for UID enrolment or Aadhaar integration most be implemented at this lowest level. However, the lack of standardization prevents tracking effectiveness of roll-out at this lowest level. Only aggregate data is available to the officials at a state or district level. For example, since the phase 1 enrolment data did not have standardized village names, it was impossible for authorities to make decisions on allocation of enrolment kits at a village level. There was no visibility as to which village had lesser enrolments as compared to other villages. Hence, it is critically important that before the SRDH data is standardized before making use as the basic identity infrastructure across applications and systems. For this, Government of Maharashtra has implemented the Usable SRDH (U-SRDH) with the objective of more effective and meaningful use of resident data.
  An example of the non-standardized data in the SRDH database is shown in the table below, where few different ways of writing state name have been depicted. Overall there were more than 700 distinct ways in which 'Maharashtra' had been written in the local language.

Table 13: Example of non-standardized data in SRDH

State Name in SRDH database (in Marathi)	Number of Occurrences



State Name in SRDH database (in Marathi)	Number of Occurrences
महाराष्ट्र	28704639
महाराष्ट्री	1543798
महराष्ट्र	160780
महाराष्ट्	104823
मी	16307
महाराष्ʼ	15062
महारष्ट्र	14974
Maharashtra	13060
महाराष्ट	8995
महारासी	8831
आहाराष्ट्र	8246
महारी	7011
महारि	5663
मी आहाराष्ट्र	3845
महा	3346
माहाराष्ट्र	2284
मई	2055
ममहाराष्ट्र	1643
महारी	1634
,महाराष्ट्र	1460
मी	1450
मुंबई	1167
महारी ाती	1058
र्मी	846

- Data de-duplication on demographic parameters in eGovernance systems and beneficiary lists: One of the key steps in the pre-processing of beneficiary lists is to find duplicate data. These duplicates in the data may be because of missing software controls or may be intentionally introduced into the beneficiary lists by malevolent entities. These duplicate records need to be investigated and removed before further processing for Aadhaar integration is performed. Some examples of potential duplicates in the beneficiary lists are as shown below:

  In the experiments conducted in the UID Innovation center, demographic de-duplication was also performed on key databases such as SRDH, election database etc. Some examples of duplicates found are as follows:

- Use of KYR details for identification of potential ‘ghost entries’ in beneficiary lists: One of the key benefits that should be realized during the seeding process is for the verification of KYR+ details of the beneficiaries. This can result in identification of bogus or ghost entries in the beneficiary lists. For example, in the list of junior college students availing benefit under scholarship schemes, KYR+ verification may show results where actual age (as recorded in SRDH) shows that the individual is a 10 year old. Such records can be investigated and verified so that only genuine entries are provided with benefits. Example of potential ghost entries are as shown below:

Table 14: Example of potential ghost entries in beneficiary lists

Name of Bank	Full Name In English	Full Address	Account Number	UID	Mobile
State Bank of India	RAHUL DILIPKUMAR TARTE	Masked – Address 1	XXX494	864XXX339	91XXX611
Bank of Baroda	RAHUL DILIPKUMAR TARTE	Masked – Address 1	XXX702	864XXX339	91XXX611
IDBI Bank	VINAYA JAGANNATH SAVANT	Masked – Address 2	XXX539	595XXX350	99XXX536
Bank of Maharashtra	VIJAYA JAGANNATH SAVANT	Masked – Address 2	XXX437	595XXX350	99XXX536
Bank of Maharashtra	VINAYA JAGANNATH SAVANT	Masked – Address 2	XXX786	595XXX350	99XXX536

Table 15: Example of KYR verification

College Name	Applicant Name	Birth Date	UID Number	Date of birth (From KYR data in SRDH)
Masked – Jr. College 1 Chandur Bazar	DIPALI VIJAYRAO INGLE	**1995-10-07**	437XXX255	**2003**
Masked – Jr. College 2 Chandur Bazar	VAISHALI HARIBHAU SURALKAR	**1984-10-07**	949XXX808	**1990**
Masked – Jr. College 3 Chandur Bazar	PRANJALI SATISH RAUT	**1996-07-05**	688XXX129	**2003**
Masked – Jr. College 4 Chandur Bazar	NEHA RAJENDRA BAND	**1996-10-13**	856XXX751	**2001**

- Verification of authenticity of bank account details for Direct Benefit Transfer: One of key requirements for Aadhaar integration, particularly for use as identifier for Direct Benefit Transfer, is that the bank account details and its mapping to a correct Aadhaar number needs to be verified. For direct benefit transfer to be effective, the right person should get the right amount in the right (or the account of choice). To ensure this, the Directorate of Information Technology conducted test / quality assurance runs where the following two steps were performed:
    - KYR+ verification to ensure that the mapping of bank account to UID is correct. Some examples of exceptions / defects encountered are as shown below:

  - After verification, INR 1/- was pushed in the bank accounts of the verified accounts to see if the transaction was successful completed. With this step, invalid bank accounts and other issues were identified for further quality assurance and control.

- Maintaining AUA ASA authentication history to reduce dependency on the Central ID repository (CIDR): Guided by Section 43A of the Information Technology Act, 2008 the Directorate of Information Technology (DIT), Government of Maharashtra is currently implementing and maintaining security practices and procedures for the identity information in the Aadhaar (UID) databases. One of the key decisions taken by the DIT, Maharashtra is that the most sensitive personal data of the residents, i.e. biometric information, will not been pulled into the SRDH. The Maharashtra SRDH contains only demographic information of the residents in the State Level database. For biometric authentication purposes, the KYR+ database maintains tables that store the history of AUA ASA authentication. The applications and systems that need authentication to be performed would be able to access this information through web services and make a decision on whether a new authentication transaction needs to be conducted or is the last transaction recent enough for the requirements of the application. For example, issuance of an affidavit or certificate through a citizen service center, it may not be required to authenticate the identity of a particular resident for every transaction. Once the first authentication is performed, that history is available and can be accessed as a reliable mechanism for verifying identity in a given period –say one week. For all subsequent transactions during the week – repeating the authentication may not be required.

8. Methodology adopted for Case Writing

This case study was been developed based on the experiments and research conducted by the UID Innovation Center set-up by Directorate of Information Technology, Government of Maharashtra. The identity information managed by the Center has been classified as "Restricted Access" in accordance to the guidelines stated in National Data Sharing and Accessibility Policy (NDSAP) and hence has been masked in the examples stated in the case study. Further details could be made available on a case-to-case basis after evaluation and approval of the Directorate of Information Technology, Government of Maharashtra.

9. Case Fact Sheet

Table 16: Case Fact Sheet

Fact #	Particulars / Parameters	Value
1	Location of UID Innovation Centre	Room No. 516, 5th Floor, Annexe, Mantralaya, Mumbai
2	Contact details	sec.it@maharashtra.gov.in
3	Total No. of Servers dedicated for UID Innovation Centre	6 Servers (through cloud implemented at Maharashtra State Data Centre): ▪ 1 Server (32GB RAM, 1 TB)

Fact #	Particulars / Parameters	Value
		▪ 1 Server (32 GB RAM, 2 TB) ▪ 1 Server (96GB RAM, 1 TB) ▪ 3 Servers (64GB RAM, 300 GB)
4	Human Resources dedicated to UID innovation Centre	1 Senior Consultant 1 Project Manager 1 Team Leader 2 Team Members (Peak Team Size: 4 Team Members)
5	Number of records in SRDH (February 2013):	42125495
6	Number of duplicates found in SRDH:	9270
7	Largest dataset on which de-duplication has been performed	Election database (7.2 Crore records)
8	Largest dataset for which seeding has been performed	Election database (7.2 Crore records) matched with U-SRDH (4.2 Crore records)