21ST CENTURY **IDENTIFICATION** SYSTEMS
DATA—POLITICS—PROTECTION

THU **NOV19** | FRI **NOV20** | SAT **NOV21** | 2015

The Dataset and The National ID Number

Panel 3: Government Data

Friday, November 20, 2015

By Deborah Rose, PhD

Visiting Scholar, FXB Center, Harvard

The Dataset and The National ID Number

The major topics in this presentation will be:

1. Civic Identity as a basic human right.
2. Three types of datasets.
3. What is a formal dataset?
4. What is a number?
5. Why propose including a Hindu-Arabic ID number field in all national identification systems?
6. What makes a good national ID number?

1. Civic Identity as a basic human right.

- Goal 16 of the UN Sustainable Development Goals (SDG) is to: “Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels.”
- Subgoal 16.9 states: “by 2030, provide legal identity for all, including birth registration”.
- In my view, every person has the right to be registered at birth, and to be assigned a national ID number. It is not one or the other, we all need both.

2. Three types of datasets.

1. Amorphous data
2. Dynamic databases
3. Formal, rectangular datasets

2.1 Amorphous Data

- Amorphous data comes from many sources.
- A dataset with amorphous data is not created all at once.
- Amorphous data may be derived from commercial transactions, web searches, buying patterns, weather patterns, etc. and is generally generated for a purpose other than analysis.
- Amorphous data does not generally target individuals.
- Such data must be aggregated before analysis.
- New techniques and algorithms have been developed to analyze it.
- The trendy term for large amorphous datasets is “Big Data”.

2.2 Dynamic Databases

- This kind of database was developed to hold administrative data. The user can search for an individual record or person, in order to confirm an obligation or benefit, but its main role is to help government agencies carry out their designated functions at the individual level.
- In the United States, at the federal level, the Social Security Administration has the authority to assign security numbers to individuals and to collect, and store basic demographic information such as birthdate, name, and sex.
- In the United States, the state is the level that is authorized to assign and maintain registries of birth and death certificates.
- In the United States, the state is the level that is authorized to issue drivers licenses with increasingly technological components, that can serve as identity documents at both the state, and, within the provisions of the Real ID act, the federal level.

2.3 Formal, rectangular datasets

In a rectangular dataset that contains information about people, a row represents a person and a column, or group of columns constitutes a variable. The format of a rectangular dataset is documented in a codebook, which lists the variable name, variable label, the values a variable can take, and the labels for each value. It may also list other attributes of the data, such as whether the variable is alphabetic or numeric.

Sources of Formal Datasets

- Federal datasets can be generated by a population census, such as the US Census, or by one of the many sample surveys fielded by the Census Bureau or other government agencies.
- University researchers may develop questionnaires or patient record forms to capture the aspects of the topic they want to study.
- Data can be extracted from a dynamic dataset with a defined range in time, person, or place, for use in analysis, according to legal constraints.
- Public-use datasets can be created, that are carefully formulated to protect the privacy of individuals, by restricting the details of location, age, race and ethnicity, income, and other demographic variables or sensitive information.

Principal Statistical Agencies

The US federal government designates 13 separate departments as “Principal Statistical Agencies” whose primary function is “the collection, analysis, and dissemination for statistical purposes. . .” The two that are most relevant to the use of a National Identification Number in creating research datasets are:

- The National Center for Health Statistics (NCHS), Department of Health and Human Services (DHHS)

- The Office Research, Evaluation, and Statistics (ORES), Social Security Administration (SSA)

They all subscribe to the “Commitment to scientific integrity, “a common set of professional standards and operational practices designed to ensure the quality, integrity, and credibility of their statistical activities.”

See the CDC website: www.cdc.gov./nchs/about/integrity.htm Accessed 19 Nov 2015

Entering Data into a Formal, Rectangular Dataset

- Formal data can be entered through special laptop interview software (CAPI), or using data entry software that includes range and value checks, algorithms that check for internal consistency, including age, sex, and answers to previous questions.

- A demographic check variable would prevent a woman from being asked about prostate cancer screening tests, or a child from being asked about the number of pregnancies she has had.

- I do not recommend manually entering data into an Excel spreadsheet as a good way to create a formal rectangular dataset. There are too many sources of error, and the file format is proprietary.

Cleaning a formal dataset

Formal datasets may be “cleaned” at the record level during data entry, and at the batch level, using extensive computational operations, which may include:

- Range checks
- Value checks
- The application of decision logic tables to account for all possible combination of answers between adjoining variables. The result is an internally consistent dataset that is ready to analyze.

Age

- Age is an interesting example of a variable that seems straight-forward to those in developed countries, but is not, for many underserved populations, including migrants, refugees and those without documents.
- A person may be asked for both date of birth and current age, or, about a series of life events if birth date is not available.
- Computing algorithms can be used to produce a simple, consistent age, where one was not given directly, for use in a dynamic database or for analysis.
- Estimation of age is needed when assigning non-contemporaneous birth certificates in countries with a universal ID mandate, such as the Aadhaar Number program of India.

Analyzing a formal dataset

- The purpose of the analysis can be descriptive, to test a hypothesis within the dataset, or to develop a model to use to estimate relationships and events in other existing datasets, or to predict the results in new datasets still to be collected.
- With the codebook, and a selection of statistical procedures, a formal rectangular dataset can be analyzed using a standard statistical package such as SAS or SPSS.
- “R” and other software packages that attempt to load the whole file into RAM cannot handle the large size of government datasets which can contain 100,000 persons or more, and thousands of variables.

Looking inside a formal dataset

Here is a simple dataset with 5 people and 7 variables:

```
369028670061 1939 07 31 2 Anderson          Barbara
147303288855 2005 09 01 1 Block              Andrew
660015251967 2015 01 26 1 Reyes              Carlos
172717215081 1960 11 19 1 Smith              John
685411833968 1945 10 15 1 White              Cynthia
```

Codebook for a formal dataset

Variable Name	Columns	Codes
IDnumber	1-11	random
CheckDigit	12	calculated
Birthyear	13-16	4 digits, 1900-present
Birthmonth	17-18	2 digits: 01=Jan, 02=Feb, 03=Mar, etc.
Birthday	19-20	2 digits: 01-31
Sex	21	1 digit: 1=Male, 2=Female, 3=Other
Lastname	22-40	Last name (up to 20 characters)
Firstname	41-60	First name (up to 20 characters)

Q. What is a merged (joined) file?

Sometimes information kept in separate files must be combined in order to carry out an analysis. Two of the major ways of combining data from different sources include:

1. A one step match-merge between two files, when the matching (ID) variable exists in both files.
2. A relational join, where the datasets do not contain a common ID, but there is a third file (a “crosswalk”) that contains both ID variables, and is used in a two-step merge.

Example of a one-step match-merge

File 1 ID Number File		File 2 Bank file		File 3 ID, Demog. and Bank Info
IDNum100 Bdate MF		IDNum100 Accounts		IDNum100 Bdate MF Accounts
IDNum101 Bdate MF		IDNum101 Accounts		IDNum101 Bdate MF Accounts
IDNum102 Bdate MF	+	IDNum105 Accounts	=	IDNum102 Bdate MF
IDNum103 Bdate MF		IDNum106 Accounts		IDNum103 Bdate MF
IDNum104 Bdate MF		IDNum108 Accounts		IDNum104 Bdate MF
IDNum105 Bdate MF		IDNum110 Accounts		IDNum105 Bdate MF Accounts
IDNum106 Bdate MF				IDNum106 Bdate MF Accounts
IDNum107 Bdate MF				IDNum107 Bdate MF
IDNum108 Bdate MF				IDNum108 Bdate MF Accounts
IDNum109 Bdate MF				IDNum109 Bdate MF
IDNum110 Bdate MF				IDNum110 Bdate MF Accounts

Example of a two-step relational merge

Part 1: The original files

File 1
ID Info File

IDNum100	Bdate MF
IDNum101	Bdate MF
IDNum102	Bdate MF
IDNum103	Bdate MF
IDNum104	Bdate MF
IDNum105	Bdate MF

File 2
Crosswalk File

IDNum100	TaxNum100
IDNum101	TaxNum101
IDNum105	TaxNum102
IDNum106	TaxNum103
IDNum108	TaxNum104
IDNum110	TaxNum105

File 3
Tax File

TaxNum100	TaxInfo
TaxNum101	TaxInfo
TaxNum102	TaxInfo
TaxNum103	TaxInfo
TaxNum104	TaxInfo
TaxNum105	TaxInfo

Example of a two-step relational merge
Part 2: Merge 1.

```
Merge 1:  ID Number File + Crosswalk File

File 1                  File 2                      File 4
ID Info File            Crosswalk File    =    ID + Crosswalk File

IDNum100 Bdate MF    IDNum100 TaxNum100      IDNum100 Bdate MF TaxNum100
IDNum101 Bdate MF    IDNum101 TaxNum101      IDNum101 Bdate MF TaxNum101
IDNum102 Bdate MF +  IDNum105 TaxNum102  =   IDNum102 Bdate MF TaxNum102
IDNum103 Bdate MF    IDNum106 TaxNum103      IDNum103 Bdate MF TaxNum103
IDNum104 Bdate MF    IDNum108 TaxNum104      IDNum104 Bdate MF TaxNum104
IDNum105 Bdate MF    IDNum110 TaxNum105      IDNum105 Bdate MF TaxNum105
```

Example of a two-step relational merge
Part 3: Merge 2.

Merge 2: ID Number File + Crosswalk File + Tax File

File 4 + File 3 = File 5

File 4 ID Info + Crosswalk File		File 3 Tax File		File 5 ID Info + Tax Info File
IDNum100 Bdate MF TaxNum100		TaxNum100 TaxInfo		IDNum100 Bdate MF TaxNum100 Tax Info
IDNum101 Bdate MF TaxNum101		TaxNum100 TaxInfo		IDNum101 Bdate MF TaxNum101 Tax Info
IDNum102 Bdate MF TaxNum102	+	TaxNum100 TaxInfo	=	IDNum102 Bdate MF TaxNum102 Tax Info
IDNum103 Bdate MF TaxNum103		TaxNum100 TaxInfo		IDNum103 Bdate MF TaxNum103 Tax Info
IDNum104 Bdate MF TaxNum104		TaxNum100 TaxInfo		IDNum104 Bdate MF TaxNum104 Tax Info
IDNum105 Bdate MF TaxNum105		TaxNum100 TaxInfo		IDNum105 Bdate MF TaxNum105 Tax Info

.

Q. Why are file merges needed?

A. The ability to merge files allows different kinds of data to be stored separately, to increase privacy and security.

What can you do with a numeric ID number?

- Keep different kinds of data in different files
- Use different ID numbers for the different kinds of files.
- Store pairs of ID numbers in a separate, transactional file.
- Merge distinct datasets to combine information only as needed, and only with the proper authority.

4. What is a number?

A short history of Hindu – Arabic numerals:

0 1 2 3 4 5 6 7 8 9

4.1 Hindu – Arabic numerals in India

As early as 500 BCE, mathematicians in India developed a system with a different symbol for each number from one to nine now generally called “Arabic numerals”, or Hindu-Arabic numerals. They spread to the Arab countries first, and only later to Europe, with the help of Leonardo Fibonacci, an Italian mathematician who brought the concept to Europe in 1202.

4.2 Numerals and the place number system

The Babylonians were the first to develop a true place number system, the position of the numerals represent such quantities as hundreds, tens and ones. This concept was absent from the Egyptian, Greek and Roman numerals of that time period.

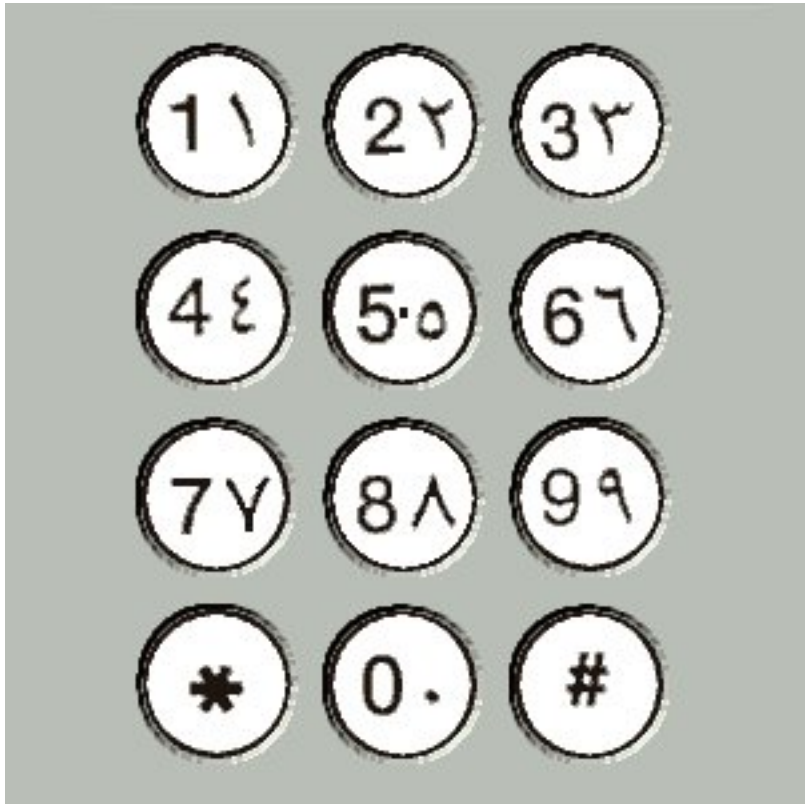
4.3 The Concept of Zero

The concept of “zero”, was developed independently by the Babylonians, the Mayans, and the Indians, first as a place holder, and then as a number in its own right, with its own properties. Division by zero was not understood until 1675, by Isaac Newton and Gottfried Wilhelm Leibniz.

4.4 Computer Codes for Hindu-Arabic Numerals

Binary Code	Hexadecimal Code	Glyph
0011 0000	30	0
0011 0001	31	1
0011 0010	32	2
0011 0011	33	3
0011 0100	34	4
0011 0101	35	5
0011 0110	36	6
0011 0111	37	7
0011 1000	38	8
0011 1001	39	9

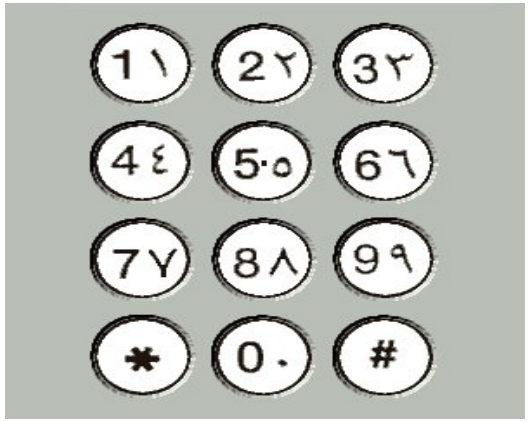
4.5 Question: What is this?



4.6. Answer: An Egyptian phone keypad

Those who read and write Arabic will recognize two forms of Arabic numerals:

- Western Arabic/European numerals (on the left)
- Eastern Arabic numerals (on the right)



5. Why propose including a Hindu-Arabic ID number field in all national identification systems?

1. It avoids the need to convert between languages with non-Roman scripts and English.
2. Arabic numerals and their ASCII / ISO codes can be read and manipulated by all programming languages and data systems.
3. It is easy to program a dataset merge and other file operations using Hindu-Arabic numerals.
4. Biometric data requires that each person have two eyes and ten fingerprints, which is not always possible, for a range of reasons. Everyone can be assigned a number.
5. Even though facial recognition images and digital fingerprints can be converted to numbers, it is simpler to use a number in the first place.

6. What makes a good National ID number?
- My personal views

1. Every country should make a National ID Number available to all residents, like the Aadhaar program in India.
2. Identification Numbers must be numeric.
3. Do not embed personally identifiable information (PII) within a National Identification Number.
4. For countries with a history of PII in the ID, use four digits for year of birth, and drop any codes for male or female.
5. Do not code race, ethnicity, or religion into the identification number, and do not put them on any identity card.
6. Where feasible, consider using different numbers for different purposes, rather than one omnibus number.
7. I champion the ID number, but am agnostic as to whether or not to require, or even issue a National Identity Card.

Final thoughts

While my own expertise focuses on the numeric and computational aspects of National ID Numbers, this conference has highlighted the great importance of the human rights component, in both historical and current use.

We must work toward a world in which national identification systems will be used only for good, and not for evil.