Long Title: Development of Metadata, Conversion and Archiving of the Time Series Data of the Completed Censuses and Surveys of the BBS.

Short Title: Historical Data Conversion and Archiving.

Mr. Chandra Shekhar Roy[1]

Abstract: *Bangladesh Bureau of Statistics (BBS) has vowed to convert Census/Surveys micro data for next generation technology format use processed in IBM proprietary technology. After Bangladesh's independence, there is a rich repository of statistical information in IBM 360 to ES/9000 mainframe tapes, dating back to late 1970s and early 2000s time.*

The overarching objective is to strengthen the prevailing national statistical archiving system. BBS will be making available of this large volume of converted data to the citizens of Bangladesh (if needed world community) so that academic and scholarly debates can take place taking cognizance of historical data.

This is a big step towards dissemination of Big Data covering Bangladesh's economic developmental trend since its birth in 1971. By revisiting time series data, it is hoped that well-informed and meticulous policies can be designed and formulated in the future. Most fundamentally, availability and easy accessibility to such a large volume of Big Data will inspire reassessing economic theories and indicators of development informing Bangladesh's position in global rankings like SDG.

Key Words: *mainframe, historical data, IBM, archiving, SDG.*

Project Name: Data Conversion, Metadata preparation, Preservation & Time series data compilation. (DCMPT)

1. Context

The Bangladesh Bureau of Statistics (BBS) is the government's apex body providing technical and administrative guidance in matters of all official statistics in Bangladesh. BBS is the national statistical organization (NSO) responsible for collecting, compiling and disseminating statistical data of all the sectors attributable to the Bangladesh economy. The underlying function is to meet data-needs of diverse users and stakeholders, e.g. national level planners, development partners and other agencies. Presently, the Bureau of Statistics possesses a large volume of statistical information dating back to the time of Bangladesh's independence since 1971.

However, much of these data sets were stored on computer media that is no longer in common use, deterring accessibility to such historical data. In 1973, BBS installed the IBM System 360, followed by IBM System 4341 in 1981 and IBM ES/9000 a year later. Following emergence of disk-system, these mainframe tapes, as a storage tool, faced serious criticism with their maintenance costs in Bangladesh. With the onset of personal computers and a proliferation of statistical software in the 1990s, the IBM ES/9000 was discontinued from 1999 onwards.

Fortunately a huge quantity of 9-track tape was used to preserve the micro data. So there is a rich repository of statistical information (census, survey and geo-code) in these mainframe tapes, dating back to late 1970s and early 2000s.

The Government of Bangladesh has initiated a project to convert all the time series statistical microdata set in a re-useable next generation technology format. In this initiative, a legacy data recovery lab will also be set up in the BBS.

[1] Senior Maintenance Engineer-IT, Bangladesh Bureau of Statistics, Government of Bangladesh, email: csroy.sme@bbs.gov.bd.

2. Towards Big Data

The overarching objective of this step is to strengthen the prevailing national statistical archiving system. The BBS will be making available of this large volume of converted data to the users in Bangladesh and for the world community, so that academic and scholarly debates can take place for the cognizance of historical data.

Data recovery and conversion from mainframe EBCDIC format into ASCII is a big step towards dissemination of Big Data covering trend of economic development in Bangladesh since her birth in 1971. By revisiting historical trends, it is hoped that well-informed and meticulous policies can be designed and formulated in the future. Most fundamentally, availability and easy accessibility to such a large volume of Big Data will inspire revisiting economic arena and indicators of development informing Bangladesh's position in global rankings.

3. Alignment with the National and Global Commitments

Along with the global commitment (SDGs), one of the strategic goals of the National Strategies for the Development of Statistics (NSDS) in Bangladesh relates to data archiving which ultimately hopes to consolidate a data disseminating and sharing system. For example,

The Sustainable Development Goal :
Goal : 9.17
    Target : 9.5, 17.6 to 17.8 (Partial)
    Backbone of data Platform, Data mining.

NSDS Strategic Goal :

Goal 3: Development of Efficient Data Management System & Maintenance.
    Indicators : 3.3
        Data management and dissemination including micro/meta data
Goal 7: Data Archive & Networking
    Indicators 7.1 :
        Legacy data archiving
Goal 8: Forward to cloud computing
    Indicators : 8.4
        Development of e-service and sharing system based on cloud computing.
        Open Government Data (OGD)

4. Objectives of the project:

The specific objectives of the project are, as follows

- To convert/rescue various census/survey data since BBS established which can be used in OGD portal of Bangladesh;
- Metadata preparation and its evaluation;
- To arrange census/survey micro data in proper time sequence;
- Preserve all these data in scientific way (future IT compatible) for reusing.
- Digital data display including 'population clock' presentation;
- Another objective is to set up a legacy data recovery lab in BBS.

5. Justification/Outcome of the Project:

- Supply older/historical data to the researcher's, stakeholders along with Metadata;
- This information will be helpful to review the history of progress of the country;
- Opportunities for sharing historical data both local and foreign countries;
- Legacy data set, will be helpful in big data preparation;
- Time series data, to predict future values based on previously observed data.

Therefore it is important to have a sound, methodological approach by which BBS can undertake Data Conversion projects, which will help to confront unpleasant surprises on later stages and resolve those issues fast and effectively.
Data Conversion project activities starts with Planning, leading to Analysis and Design, progressing to Conversion of Data, finishing in Metadata - where converted data loaded in the Time series database.
In this stage BBS, in conjunction with IT expert & Statistician, prepares a plan, which is also intended to shorten the duration of the Conversion along with metadata process and also reduces business impact and risk.

6. Challenge of the project:

- Total 8600 the huge numbers of 9-track spool data tapes needs to be read and convert.
- Technology outdated
- More than 20 years old backup tape
- Preservation environment was vulnerable
- Some report/publication still unavailable for metadata preparation
- Historical data up to 1999 to present data from 2000 to 2019 needs to be developed time series data.

Other Factors contribute to complexity of such projects are

- In-depth understanding of the functionality of the source data structure.
- Source data quality, if poor, needs to be cleansed in order to be successfully converted.

7. Success/probability/prospect of the project:

- Multiple backup tapes for each census/survey;
- IBM mainframe backward compatible technology exist;
- EBCDIC to ASCII conversion software is available in USA market;
- Magnetic tape life expectancy 10-30 years (*Dr. John W. C. Van Bogart,* National Media Laboratory, USA)
- Some tapes contain COBOL program and data layout;
- All the backup data was stored in structured way (from 1972 to 1999);
- Experts on mainframe data are agreed to help us;
- Legacy statistical publications can be found in other organizations of the country;
- Recovered micro data set will be preserved using new technology in earth quick free zone;
- International research community may help us for getting legacy data;
- Our project has been piloted in the University of Minnesota.

8. Benefits /Future vision of the project:

Data community in Bangladesh is realizing it as ‘Digital Assets’ of the country.
This project can be shared in the international community both data rescue, Big data and dissemination groups.
SDMX can be implemented easily by using all those data along with current data.
SDMX can be initiated in BBS where 47 years of data are available.

Bangladesh is going to cross the LDC boundary so many of our Government organization may need to convert their legacy data to the modern era. BBS data recovery lab can be the pioneer in this regards.
Data preservation is a vital issue in a developing country. Special attention may need to support in this data-security field.

9. Key Strategies of Successful Data Conversion:

The complexity of data conversion requires that certain strategies be put in place. Here are some essential strategies that we must apply in order to ensure the success of our data conversion projects.

A. Proper Planning:

Any data conversion project needs to start with defining the boundaries of the project. These include:
- What kind of data needs to be converted?
- What is the Source of data and format?
- What is the quality of data and its availability? Does it require full or partial conversion?
- What kinds of formats are needed for data conversion?
- What would be the tentative duration of project?

B. Correct Execution:

In our dram project 60% percent works are Machine intervention i.e. re-engineering where a group of IBM mainframe compatible hardware(9-track tape drive, Spool tape cleaner) need to be used. More than one scientific oven also need to be used to remove fungus and stickiness of 9-track spool tapes.

i. Exclusive data conversion software need to be used to convert data from EBCDIC to ASCII format. a) First of all this software will capture the data from the magnetic and will create an Image file, So that physical tapes no more needed in future. b) Another beauty of this software is It will create ASCII text from the captured image files. Finally, we will get the numerals data i.e. 0 to 9. Statistically we called it data records. All these records are meaningless unless we shape it with proper layout/dictionary.

ii. Rest 40% of our project work is Human intervention. Within the 8600 tapes there is programming tapes, data tapes and geo-code tapes. Most of the programs were written under COBOL language. In the programming and geo code tapes the output comes numerals with characters. There is a challenge if programming tapes fails to read or convert corresponding data tapes, we need third party statisticians or SME assistance. Defining and implementing data quality standards to ensure consistency across the different databases.

   a) Their duty is to create metadata of the converted ASCII file data. In this case, they need hard copy of survey/census data collection questionnaire/schedule. It can be found in the publication of that particular census/survey.

   b) Data Profiling and Cleansing: Ensure that proper data profiling and data cleansing procedures are in place so that the original data is of high quality. This helps to smoothen out the subsequent data conversion procedures.

   In light of data users view metadata should be very constrictive and under a database also, we called it Metabase (metadata + database) where each and every instructions need to be mentioned. Following data conversion, ensure that the duplicate master data has eliminated, reducing the risk of incorrect transactions and unreliable reports. The project should satisfy all principles of data management and data governance. The project adherence to FAIR principles: The DCMPT philosophy has always been consistent with the principles of Findability, Accessibility, Interoperability, and Re-usability.

iii. Third objective of our project is to preserve all those micro data in a scientific way so that stockholders/data journalist can use it future technology format.

iv. The final goal of our project is to create or develop time series data both census and surveys from year1972 to 2021. Data conversion is thus, a task critical from both business and technical perspectives.

v. Documentation of Legacy Fields: Each data element in the legacy system should have defined Meta data as well as documented valid values that apply to specific periods. For instance, if field "X" was used from 1974 - 1980 as a business element; it was defined one way. Then in the 1980s field, "X" was no longer needed because of a business decision. Collecting Meta data information is tedious and time-consuming; however, it will make the legacy data conversion much easier saving time, money and resources.

Annex I: Project Summary:

Project title:	Data Conversion, Metadata preparation, Preservation & Time series data compilation (DCMPT)
Sponsoring ministry/ Division:	Ministry of Planning, Statistics and Informatics Division
Executing agency:	Bangladesh Bureau of Statistics (BBS)
Location of the Project:	The location of the project in Dhaka, Bangladesh
Project period:	July 2017 to June 2019
Number of backup tapes:	8600 nine-track spool tapes (Reject tape, duplicate tape & data tape)
Series of census:	24 censuses (including 3 decennial censuses)
Series/Types of surveys:	46 surveys (HES, DHS, LFS, CPS, AES,MFG, MICS, LOS, etc.)
Cost of the Project:	80.60 million Bangladesh Taka
Source of Fund:	Full GoB (Government of Bangladesh)
Resource manpower:	From Bangladesh (Present & retired BBS IT experts & statisticians)

Annex II : Profile of the Project Director

Mr. Chandra Shekhar Roy, Senior Maintenance Engineer-IT. Working in Bangladesh Bureau of Statistics (BBS), the NSO of this country. I have been graduated in EEE and master degree in Computer Science. The Government of Bangladesh has started microdata related project where the data need to be converted to reusable format since independence in 1972. This data revolution will improve our knowledge and benefit our society through data-driven research and innovation. I am the project director of this dream project.

I was involved in Paris21s IDR project, iData Studio project, MPC, ANCSDAAP, ITU & OGD focal point officer in BBS. Please read the brief information related to my project *(.pdf attached).*

Contact details:

Mr. C. S. Roy
Senior Maintenance Engineer-IT,
Project Director, DCMPT project,
Bangladesh Bureau of Statistics, Dhaka.
email: csroy.sme@bbs.gov.bd
cell: 8801556462036
skype: cs_roy1