Report

# of the

Committee on Online Reporting Systems

सत्यमेव जयते

GOVERNMENT OF INDIA

NATIONAL STATISTICAL COMMISSION

MINISTRY OF STATISTICS AND PROGRAMME IMPLEMENTATION

NEW DELHI

Request for comments and suggestions on the Reports of Committees constituted by NSC

NSCs vide letter No.8(64)/2010-NSC dated 05.10.2016 constituted 5 professional committees to examine potential improvement in methodology and database related issues pertaining to estimation of GDP within the broad framework of SNA 2008. These committees are 1) Committee on Real Sector Statistics, 2) Committee on Financial Sector Statistics, 3) Committee on Fiscal Statistics, 4) Committee on Online Reporting System and 5) Committee on Analytics.

Four of the above committees submitted their reports in the 102nd Meeting of the National Statistical Commission held on 16-17th July, 2018 under the Chairmanship of Dr R B Barman, Chairperson, NSC at Sardar Patel Bhawan.

The draft reports of these committees are placed in the public domain to facilitate wider public consultation. The NSC welcomes comments and suggestions on the reports at the following address by 30th September, 2018:

National Statistical Commission
NSC Secretariat, Room No-305,
3rd Floor, C-Wing, Pushpa Bhawan, New Delhi – 110062
Email: nsc-secretariat@gov.in

It may be noted that the NSC does not necessarily agree with the views, data and other contents of the reports.

CONTENTS

Acknowledgements

As Chairman of the Committee on Online Reporting System, I had the privilege to have some of the distinguished personalities of the field as member of the Committee. Their valuable inputs in discussions and contributions helped me to shape the report for which I am thankful of them.

I gratefully acknowledge guidance received from Dr. R.B. Barman, Chairman of the National Statistical Commission at every stage of preparation of this report.
I would also like to acknowledge the support of Shri Rajeev Lochan, DG (SS) who helped us in providing the inputs regarding data collection system in the field of Agriculture.

My thanks also go to Shri Panchanan Dash, ex-ADG, DSDD, Shri Pravin Srivastava, ADG, NAD in successfully convening the meetings of the Committee. My special thanks go to Shri B.N. Tiwari, ADG (DSDD) for giving final shape to the draft report and co-ordinating with all the Members of the Committee and to Mr. Manoj Kumar Gupta (Director, DSDD) for preparing minutes of meetings of the committee. I would also like to thank the NSC Secretariat without whose support, it would not have been possible for the Committee to conduct its meetings successfully.

I would also like to acknowledge the support of FOD, NSSO in arranging the meeting at Pune, Maharashtra and Ahmedabad, Gujarat.

The Committee had three formal meetings at Mumbai, New Delhi and Pune.

Dr. Ashok Nag
Chairman

F. No. 8(64)/2010-NSC
Government of India
Ministry of Statistics & Programme Implementation
National Statistical Commission Secretariat

2nd Floor, Sardar Patel Bhavan,
Sansad Marg, New Delhi
Dated 5th October 2016

ORDER

The issue of constituting professional Committees to assist it on various technical issues was under the consideration of the National Statistical Commission (NSC) for quite some time. It has now been decided by the NSC to constitute five professional Committees. The details of composition and terms of reference of each Committee are given in the Annexure.

2. These Committees are expected to cover the requirement of statistics for estimation of GDP as per SNA 2008 and para 1.59 thereon, data governance for quality, timeliness and credibility of collected data and derived estimates, provide for data integrity and audit trials of a National Statistical System and the state-of-the-art system for management of data taking full advantage of Information Technology, distributed network and cloud. The system is also expected to make data and estimates available at disaggregated level by various dimensions of industry, geography, time, size class etc. as possibility, thereby throwing up deep insight for policy and evaluation thereof based on actual observations.

3. All the professional Committees will have tenure of one year initially. NSC will oversee the progress made by each committee on a quarterly basis. Each committee may co-opt member(s) on need basis, subject to the approval of NSC.

4. The non-official members will be entitled for a sitting fee of Rs. 1000/- per day for attending the Meetings of the Committees. The TA/DA entitlements and cost of local travel for the non-official members will be at par with a Joint Secretary to the Central Government for attending the meetings of the Committees.

5. Secretariat support to the Committees would be provided by the respective offices where the members Secretaries of the committees are working. The expenditure on the Committees would be met from the non-plan budget of the Ministry of Statistics & Programme Implementation (MoSPI) allocated to the NSC.

6. This sanction issues with the approval of Chief Accounting Authority. The advice of Internal Finance was conveyed vide Dy. No. 1113/AS & FA dated 04.10.2016.

7. This order comes into immediate effect.

(Sd/-)
(Deepak Mehra)
Director, NSCS

Terms of Reference of the Committee on Online Reporting System

1. To review the existing system for collection of Core Official Statistics .e.g. IIP, CPI,WPI,ASI, Consumer Expenditure Survey, Employment and Unemployment Survey etc. with a view to suggest measures for an online system capturing granular data on output , input, price , expenditure , employment etc.
2. To recommend suitable measures for automated online collection of these statistics with possibly for improving quality and timeliness
3. To suggest development of templates for collection of granular data online from primary sources and recommend way(s) to create infrastructure for their deployment

Constitution of the Committee on Online Reporting System

Dr. Ashok Nag Adviser Reserve Bank of India (retd.)	Chairman
Prof. Abhiman Das Indian Institute of Management Ahmedabad	Non-official member
Representative from Ministry of Corporate Affairs	Member
Representative from Ministry of Finance	Member
Representative from NAD, Central Statistical Office	Member
Representative from National Informatics Centre	Member
Representative from Ministry of	Member
Representative from Labour Bureau	Member
Representative from Department of Industrial Policy and Promotion	Member
Additional Director General ,FOD, NSSO	Member
Representative from Department of Statistics and information Management (DSIM), RBI	Member
General Manager , IT , SEBI	Member
CGM In-charge, IT, NABARD	Member
Representative from Ministry of Railway	Member
Director, DES, Andhra Pradesh	Member
Director, DES, Punjab	Member
ADG, Computer Centre	Member Secretary

List of Abbreviations

ABS –	Australian Bureau of Statistics
AIDIS –	All India Debt and Investment Survey
ARTS –	Annual Retail Trade Survey
ASI –	Annual Survey of Industries
ASI –	Annual Survey of Industries
CANSIM –	Canadian Socio-Economic Information Management System
CAPI -	Computer – Assisted personal Interviewing
CARI -	Central Agriculture Research Institute
CATI -	Computer Assisted Telephone Interviewing
CCS –	Cost of Cultivation Studies
CES –	Conference of European Statisticians
CORS –	Committee on Online Reporting Systems
CPI –	Consumer Price Index
CSO –	Central Statistics Office
CSV –	Comma Separated Values
CURFs –	Confidentialised Unit Record Files
DE –	Directory Establishments
DES –	Department of Economic & Statistics
DESAg –	Directorate of Economics & Statistics, Ministry of Agriculture
ELECTRA -	Electrical , Electronics and Communications Trade Association
EPFO –	Employees Provident Fund organisation
Eurostat-	European Statistical System
FSI-	Forest Survey of India
GDP-	Gross Domestic Product
GSBPM –	Generic Statistical Business Process Model
GST –	Goods & Services Tax
GVA –	Gross value added
IBM –	Indian Bureau of Mines

IIP -	Index of Industrial Production
IISc –	Indian Institute of Science
IIT -	Indian Institute of Technology
IMDB –	Integrated Metadata Database
IRDA -	Insurance regulatory and Development Authority
ISI –	Institute for Scientific Information
KMZ –	Keyhole Markup language Zipped
MRR –	Metadata Repository & Registry
MRTS –	Monthly Retail Trade Survey
NABARD –	National Bank for Agriculture and Rural Development
NAD –	National Accounts Division
NAS-SM –	National Accounts Statistics: Sources and Methods 2012
NBFC –	Non Banking Financial Company
NDE –	Non Directory Establishments
NIA –	National Income Accounting
NIC –	National Informatics Centre
NSC-	National Statistical Commission
NSSO –	National Sample Survey Office
OECD –	Organisation for Economic Co-operation and Development
OLAP –	Online Analytical Processing
OLTP –	Online Transaction Processing
ONS –	Office of National Statistics
ORS –	Online Reporting Systems
OSS	Official Statistical System
P & L –	Profit & Loss
PAPI –	Paper and Pencil Interview
PDF –	Picture Document Format
PFRDA –	Pension fund Regulatory Development Authority
RADL –	Remote Access Data Laboratory
RBI –	Reserve Bank of India

RDD –	**Random Digit Dialing**
SAP -	**Systems, Applications and Products**
SDMX –	**Statistical Data and Metadata eXchange**
SEBI –	**Security and Exchange Board of India**
SFDs –	**State Fishery Departments**
SN –	**Statistics Netherlands**
SNA -	**System of National Accounts**
STATCAN –	**Statistics Canada**
UNECE –	**United Nations Economic Commission for Europe**
VPA –	**Visitor Pattern Analysis**
WES –	**Website Evaluation Survey**
WPI –	**Wholesale Price Index**
XBRL –	**eXtensible Business Reporting Language**

Executive Summary

Production of statistical outputs by official statistical agencies is a generic process that is amenable to standardization across national jurisdictions. Such standardization is necessary as official statistics is the principal source of information for comparison of economic performance of nations as well as of the status of wellbeing of populations of different countries. Such a standard is an evolving process and the current internationally accepted standard is known as the Generic Statistical Business Process Model or GSBPM. Data collection or data reporting is a critical, if not the most important, component of this process as it impinges on the quality, creditability and timeliness of official statistics. Formation of this committee by the National Statistical Commission is a testimony to the importance that the commission attaches to this component. Building an Online Reporting System is a modernization effort towards making official statistics achieve its stated objective.

A review of the current status of data flow to the CSO, the producer of the national accounts in India, reveals that a large part of this data flow is paper based and lacks automation at various stages of the production process. An online reporting system for collection of primary data can impose data validation rules at the time of collection itself. If such validated data can be accessed by the CSO in a format devised by it, it would help CSO significantly to achieve the required level of quality in outputs it produces. The committee tried to get the current status of data collection process in regard to administrative and other statistical data obtaining in various ministries and government departments but could not succeed. Not availability of such metadata in a readily queryable database could be the reason for this failure.

A review of the best international practices in this regard reveals that statistical agencies in most of the developed countries have made the transition from manual process of data collection and file based data storage system to online reporting system and database oriented data storage system. The committee noted that the technology of survey based data collection process has progressed in the last few decades because of enormous development in communication and computing technology in recent times. The committee believes that Indian official statistical system is capable of adopting the best international practices in this regard.

The committee has also looked into the best practices of data processing and data dissemination system that is a perquisite for creating a technology driven official statistical system that can create trust in the official statistical outputs. Two major prerequisites are: building a proper metadata database and creating a statistical database at the CSO level. Once these perquisites are put into place, it should be possible to create a data dissemination architecture that would allow policy makers and general public to use official statistics for their use according to their requirements.

The committee strongly recommends that data should be captured in machine readable format, thus completely eliminating data transcription from paper format to electronic format. Computer assisted survey data collection process should be the norm and not exceptions. Data flow between statistical agencies should be through electronic network. An online reporting system would be of limited value if not complemented with storing of data in a proper database and all associated metadata in a query enabled metadata database.

Background

The National Statistical Commission (NSC) decided to constitute five professional committees to assist it on various technical issues (vide its order F.No.8 (64)2010-NSC dated 5th October 2016). These committees are required to examine all data related issues pertaining to estimation of GDP within the broad framework of SNA 2008. One of the most important data related issues that all the five committees are required is to examine the feasibility of building macroeconomic accounts of a sector or of the total economy by suitably aggregating unit level accounts. In this regard a reference has been made to the paragraph 1.59 of the SNA 2008, which has pointed out the imperative need for establishing such a linkage between macro and micro level datasets to ensure highest level of quality, timeliness and creditability of official statistics. The committees also need to examine the current status of use of information technology at various levels of statistical data management- from collection of data to dissemination of final data products- and recommend the best practices in this regard. The primary task of the Committee on Online Reporting System (CORS) is to examine the best practices in the area of collection/ collation of data using the latest state-of-the art technology. The committee is also required to examine and make recommendation about modes of dissemination of official statistical products.

Terms of Reference of the Committee on Online Reporting System

1. To review the existing system for collection of Core Official Statistics .e.g. IIP, CPI,WPI,ASI, Consumer Expenditure Survey, Employment and Unemployment Survey etc. with a view to suggest measures for an online system capturing granular data on output , input, price , expenditure , employment etc.
2. To recommend suitable measures for automated online collection of these statistics with possibly for improving quality and timeliness
3. To suggest development of templates for collection of granular data online from primary sources and recommend way(s) to create infrastructure for their deployment

The committee held 3 meetings on 7th April 2017, 7th December 2017 and 26th March 2018. The draft report was discussed in the 3rd meeting and its broad outline was accepted for finalization by the chairman in consultation with the member secretary.

Methodology / Approach

The committee decided to collect "as-is" position of data collection process obtaining in various government agencies entrusted with collection of basic data that are essential inputs to estimation of GDP and its components. The statistical activities of these agencies are generally carried out under the supervision of cadres of Indian Statistical Service. A questionnaire was circulated to all these agencies. The format of this survey questionnaire is given in **Annexure II**. As none of the agencies responded to the survey questionnaire, this report is based on information available from websites of various national official statistical offices. However, many of the committee members provided information about data management practices obtaining in their respective agencies in meetings of the committee.

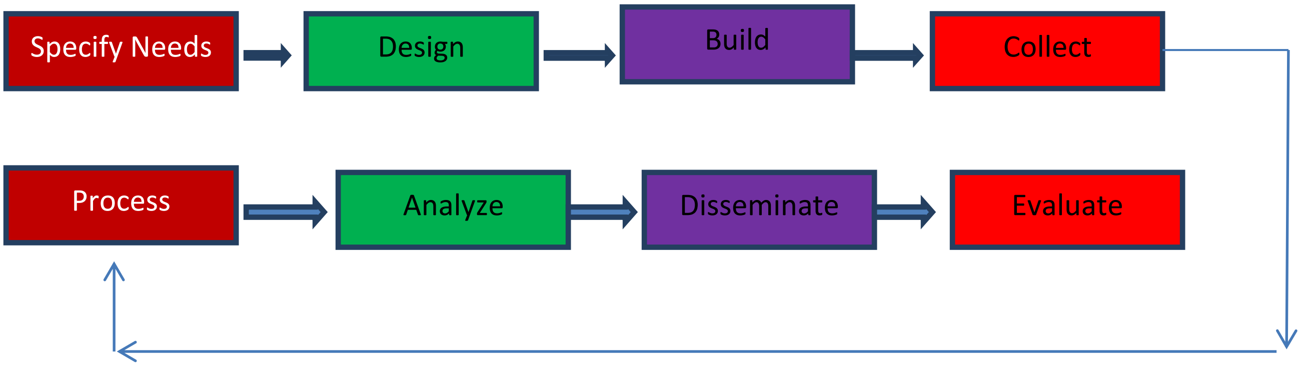
In regards to the best international practices, the websites of various national statistical agencies, United Nations Statistical Commission and Eurostat were consulted. The **Conference of European Statisticians (CES)** is a collaborative effort on the part of Chief Statisticians of more than 60 countries and a number of international organizations under the auspices of the Statistical Division of the United Nations Economic Commission for Europe (UNECE). CES organized a seminar on New Frontiers in Statistical Data Collection in November 2012. The secretariat of CES conducted a survey on Online Data Collection amongst the participants of the conference and a report summarizing responses to this survey was presented at the aforesaid conference. This report as well as the papers submitted by representatives of various statistical agencies was studied to understand the current practices of data collection process in many developed countries and its future direction .

Although the committee’s Terms of Reference did not explicitly mention certain stages of data management lifecycle that have a bearing on modes of data flow to the national statistical offices entrusted with the task of compilation of national income accounts, the committee decided to touch upon them to the extent relevant and necessary for the assigned task . These are:

a. Administrative data as a source of data for national income accounting and issues germane to its use in building macroeconomic accounts.
b. Use of Big Data in creation of official statistical products
c. Data dissemination modes that allow users to create user defined report
d. Building and maintaining a statistical database at CSO
e. Building and maintaining a metadata repository

Introduction

A national statistical organization produces statistical products for use of policy makers, researchers and finally for the citizen at large. Most of the national statistical organizations are entrusted with the task of estimation of GDP and they do it within an internationally agreed framework known as the System of National Accounts or SNA. This framework has evolved over time. The current version is known as SNA 2008. A common accounting framework also calls for a standardization of production practices, encapsulating the best practices gleaned from various national statistical agencies. Various international multilateral organizations like UNECE / Eurostat / OECD are engaged in preparing such a standard for the last decade or so. This standard is known as Generic Statistical Business Process Model (GSBPM), its latest version being GSBPM 5.0. The stages in this business process model are presented in the following diagram:



The model described above binds together all stages to meet a specific need for a statistics at its most granular level. For example, let us consider the need for estimation of GDP from Agriculture. This need is described at a very high level of aggregation. We need to break it down to the lowest level of granularity at which data can be collected and, processed and analyzed. For example, we may like to estimate output and input for a specified crop, say rice. The reporting system required for the inputs and outputs must reckon with the feasible level of granularity and collect data accordingly. In other words, it would be wrong to describe the contour of a reporting system in a generic way. It must be grounded on the specific needs, analytical requirements associated with a specific statistical output and dissemination thereof. From this perspective, the committee's effort to embed discussion on reporting system on a broader canvas of GSBPM is justified and required.

Definitions

Statistical data: - Data inputs that are used to create statistical products by statistical agencies like CSO, NSSO, Labour Bureau, DESAg (Directorate of Economics and Statistics,Ministry of Agriculture) etc.

Online Reporting System (ORS): It is an electronic platform enabling reporting entities to submit information in a structured format to the receiving agency. Surveys conducted via email in the form of an attachment or within the message body are considered as ORS but of rudimentary type. Ideally data received through ORS should automatically get stored in a database. This generally does not happen in case of data received through email attachment. Web surveys are considered as ORS and these surveys are self-administered. When survey data are collected through computing devices by interviewers and then uploaded into a centralized system, such data capture method is largely similar to that of an ORS. The main difference lies in administration of data collection process. While ORS is self-administered by the respondent, computer-assisted survey data collection is interviewer administered.

Administrative data: - This set of information is collected mainly for administrative purposes. Such data gets generated as a derivative of routine delivery of services by government agencies or for keeping record of objects like business entities, households, individuals or transactions. For example details of companies registered under companies act are maintained by the Registrar of Companies and the same can be used by the CSO for national accounting purpose.

Statistical Survey: - Ordinarily the term survey is used synonymously with sample survey. But for national statistical authorities this term encompasses mode of collection of statistical data. Censuses, Sample Survey, collection of data from administrative records are examples of various types of statistical surveys.

Register: - Register is a complete list of all objects in a given population of these objects. An Administrative register maintains list of all objects to be administered by a government authority or an administrative arm of a government.

Input Data for National Income Accounting:

“National Accounts Statistics: sources and Methods 2012*”(NAS-SM)-* a CSO publication -gives details of sources for each data item that is an input to national income measures. The publication does not give a systematic presentation of source/item/transformation used/ unit/ granularity/ format of data availability. In other words, a systematic and query enabled metadata is currently unavailable to assess the current status of input data used in national income accounting. Be that as it may, we present below a tentative summary of the sector wise source metadata available in this document.

Sector: Agriculture and allied Activities

<u>Agriculture proper</u>: Directorate of Economics and Statistics, Agriculture DESAg, is the principal administrative department of the central government providing data on agriculture for NIA. DESAg publishes all major agricultural statistics – land use, area and outturn of major and minor crops -in their annual publication- “Agricultural Statistics in India”. It appears NAD of CSO collects data on this sector from the publications. In other words the administrative data collected by State Revenue departments are not shared in electronic form with NAD. Other publications used by NAD and released by DESAg are also paper based and granularity of published data is not in the lowest available level.

For input costs, data are sourced from Cost of Cultivation Studies (CCS), presumably, as published data. Mention has been made about other sources of data like Central Electricity Authority, Fertilizer Association of India etc. Apparently such data are sourced from regular publications of these agencies. Data are also sourced from various commodity boards and commodity specific development boards. CSO does not give any further details or metadata about data sourced from these agencies.

Livestock: A nationwide survey (Integrated Sample Survey) conducted by state Animal Husbandry on a regular basis is the primary source of production data. At all India level data are consolidated and published annually by the Department of Animal Husbandry and Dairying, Ministry of agriculture. Only published data are used by NAD. Another main source of data is Livestock Census, conducted quinquennially. NAD only uses published data.

Forestry and Logging

Forestry: Product and prices data for industrial wood and minor forest products are captured by State Forest Departments. NSSO five yearly Consumption Expenditure Survey (CES) collects data on fuelwood consumption. Forest Survey of India (FSI) is also source of data for some of the products of this sector.

Fishing: The sources of data for fisheries statistics is the State Fishery Departments (SFDs). Estimates of production from coastal marine sector are compiled based on data collected through a sample survey in all the maritime states. For inland production of fishes another survey is conducted.

Sector: Mining and Quarrying

For estimation of outputs minerals are divided into two groups – major minerals and minor minerals. For Coal India, balance sheet and P&L account of Coal India is the data source. For the private sector year wise estimates published by the Coal Controller of India are used. For Petroleum and Natural Gas, the concerned ministry’s publication is the data source.

Sector: Manufacturing

The manufacturing sector comprises two broad sub-sectors, namely “Registered” and “Un-registered”. The former sub-sector comprises all establishments engaged in manufacturing of goods and registered under Indian Factories Act. The residual “”unregistered” sub-sector includes allother establishmentsengaged in manufacturing activities. The data source for inputs and outputs for “registered” sector is the Annual Survey of Industries (ASI). For constant prices estimates, WPI data are used. The various relevant aspects of ASI and WPI are described later in this report.

The “unregistered” sub-sector is again classified into two groups, namely, (i) Directory Establishments (DE) and non-Directory establishments (NDE). The data sources used for computation of GVA for this sub-sector are: All India census of Micro, Small and Medium enterprises; NSSO surveys. According to NAS-SM, NAD uses only published data. It is not known to what extent NAD has access to unit records of NSSO surveys it requires for NIA.

Sector: Construction

This sector is again classified into two broad groups- Accounted (Pucca construction) and Unaccounted (Kuccha construction). The GVAs from these two sub-sectors are compiled through a combination of commodity-flow approach and expenditure approach. Given the indirect nature of both these approaches, data sources used for NIA are diverse and many. Within the Accounted group, estimates for public and private sector are compiled based on budget data, balance sheets and P&L accounts. For the unaccounted group, apart from the above ones, AIDIS survey of NSSO is another data source.

Sector: Electricity, Gas and Water Supply

This sector is dominated by organized public and private sector. Budget documents of central, state and union territories, balance sheets and profit and loss accounts of producing enterprises are the data sources for compilation of GVA from this sector.

Sector: Trade, Hotel and Restaurants

Trade : For the organized component of this sector, balance sheets and profit and loss accounts are the main data sources. For the organized public sector budget documents are also used. Other publications used as data sources are: RBI study on the finances of companies engaged in activities of this sector; NABARD publication on Co-operative sector.

Hotels and Restaurants: For public sector enterprises in this sub-sector annual reports and Profit &Loss accounts are the data sources. For the private sector RBI study is the main data source. Since the year 2013-14 onwards RBI has started using MCA data and the coverage of RBI study is 30 percent of population PUC of Public Limited companies. RBI study covered 313 companies under Accommodation and Food Service Activities industry group which was published on RBI website.

Sector: Transport, Storage and communication

Transport by Railways- Publications such as Annual Report &Accounts and Annual Statistical Statements and Central Government budget are the main data sources.

Transport other than Railways- Apart from Annual accounts, P&L accounts of institutional components of this sector, budget documents of central and state governments are data sources for this sub-sector. NSSO surveys, Enterprise surveys are also used for estimating inputs that are required for GVA estimation.

For storage and communication data sources are on similar lines.

Sector: Banking and Insurance

This highly regulated sector is best placed in terms of having balance sheet and Profit & Loss data for the organized sector. For the unorganized sector only a ratio of the organized sector is used.

Sector: Real Estate, Ownership of Dwellings and Business Services

Except for software services, GVA for this sector is compiled using labor input and value added per worker. Thus main data sources are NSSO surveys and Enterprise surveys. For software companies, annual accounts and P&L accounts are the main data sources.

Sector: Public administration and Defense

Budgets of administrative authorities at various levels, combined with reports by Comptroller and auditor General of India are the main data sources for estimating GVA from this sector.

Sector: Other Services

For public sector organizations engaged in activities covered by this sector, standard data sources like budget documents and annual reports of the concerned units are available. For organized as well as unorganized private sector, the indirect method of GVA per worker multiplied by an estimate of workforce is used to compile GVA for a benchmark years. For subsequent years the benchmark estimates are taken forward by suitable volume and price indicator. Mostly survey data are used for this purpose.

Committee's observations on Data Sources used in NIA:

It is seen from above that a significant amount of administrative data are consumed for compilation of National Income Accounts. However, NAD sources these data only from published reports of concerned   various government departmentsand agencies. This form of data collection from published report has adverse consequences as described below and alternative mechanism for establishing an online reporting system can be worked out.

1. Data transcription from publications to worksheets of NAD is fraught with the risk of data entry errors. It also creates an avoidable time lag between data processed in the department receiving data from its originating point and publication of the same. For example, DESAg supplies Land Utilization statistics with a time lag of 1 to 2 years. If the state revenue departments are connected

directly with NAD such a time lag can be avoided. The estimates of production of plantation crops are collected from respective boards which collect data submitted by growers in the form of returns. It would not be difficult to redirect these returns to NAD which can make estimates by itself. Similarly prices data collected by state DES can flow to CSO directly as and when data are obtained by DES from Agricultural Produce Market Committees. Direct data flow through Web services or through FTP servers can be worked out for data capturing agencies like State Forest Departments (SFDs), State Fisheries Departments, Indian Bureau of Mines(IBM), Coal Controller of India, State Geological Departments for minor minerals etc. The main idea should be that, as and when data are prepared for publication by any agency, the same should be made electronically available to NAD without any time lag. NAD can prepare a specific return format for such data submission by the respective agencies and obtain it over secured network. The data should be validated at the time of data submission itself and stored in a proper statistical database.

2. A national data grid can be established for all data flow from government agencies to NAD.

3. NAD must build a statistical database some details of which are described in the section 2 of this report below.

Review of Reporting System for important datasets on prices and productions and registers.

<u>ASI:</u> Annual Survey of Industries (ASI) is the principal source of industrial statistics in India. The survey collects data from all organized establishments engaged in activities relating to manufacturing processes, repair services, gas and water supply and cold storage. The concerned respondent units are statutorily to submit the returns along with the balance sheet and other relevant documents within the prescribed period. Surprisingly returns are not submitted electronically. The duly filled in returns for the Central Sample are compiled and transmitted online to CSO (IS Wing) through ASI Web Portal for processing and publication of results. Obviously this is not a self-administered survey.

This survey can be much more effectively conducted by deploying a proper online self-administered survey. The ASI field staff can review and carry out error correction online. The large establishments should be allowed to create the ASI forms directly from their ERP systems. In fact, SAP can be easily configured to provide a substantial amount of data that ASI schedule needs. The CSO can work with SAP to create necessary reports for submission to ASI portal automatically. For authentication purpose, there are number of technologies that CSO can deploy. Notable among them is Blockchain. A Blockchain can be created for each company with timestamp and digital signature appended to it. . All relevant documents which are already submitted to some governmental authorities can be submitted in electronic form as an attachment. For example if corporate balance sheet is required the same is being submitted in XBRL form and the same can be automatically converted into a structured format that enables data manipulation in a computer based system

<u>CPI- Consumer price Index (base year 2012</u>): The National Informatics Centre (NIC) has developed two independent web portals – one for rural CPI and another for urban CPI- for online transmission of retail price data. It is not clear why the same data is not opened up to general public to receive comments on the quality of data. There cannot be any privacy issue if identity of data provider is not disclosed. Opening of such data to public scrutiny will go a long way to improve data quality.

WPI- Wholesale Price Index (Base 2011-12): The Office of Economic Adviser has developed an online data transmission mechanism for direct inputting of price data by the selected factories/ manufacturing units. It appears that an arrangement has been made with NSSO to deploy their field staff to assist the business units in submitting data and validating the same at field level. Apparently there is no built in validation process to allow submission of validated data only. NAD should have machine –to-machine connectivity to WPI database existing in the office of Economic Adviser and the relevant data should be made available through a computer mediated query system to compiler of required deflators within NAD

IIP-Index of Industrial Production (base year 2011-12): There are 16 Government of India Ministries/ Departments/ organizations which provide primary data required for compilation of IIP. Out of these only 3 sources collect data through web portal. The metadata document for IIP released by CSO claims that the line ministries responsible for collecting primary data "follow a defined protocol to transfer the data to MoSPI"; there is no information available in the public domain about this protocol. Despite enabling recommendation of the working group for new IIP for establishing a web based mechanism for primary data collection, not much progress has been made in this direction.

NSSO Surveys: All NSSO surveys follow PAPI methodology for data collection. Huge effort that is required for transcription of voluminous data leads to considerable delay in publication of results. There is no usage of linking of diverse administrative records like Census / revenue / PDS etc. to create a much better sampling frame than is the case now. To give an example, The Debt and Investment Survey must take into account commercial / co-operative/ NBFC/ Chit-fund / Insurance database available with concerned regulators. Finally, NSSO badly needs to migrate to

CAPI/CARI mediated survey methodology to produce more reliable and quicker estimates.

Initiatives taken by other jurisdictions for Electronic Data Collection:

New technologies like Big Data Analytics, Artificial Intelligence based data editing techniques, pervasiveness of Internet; Computer Aided data capture method etc. are revolutionizing the way official statistical system used to function in various countries. Considering the importance of the topic, Conference of European Statisticians under the auspices of United Nations Economic Commission on Europe (UNECE) organized a conference on New Frontiers in Statistical Data Collection in November 2012. Many participating countries presented the initiatives taken by their respective national statistical agencies in this area. The second seminar on Statistical Data collection was held in Geneva, from in September 2013. The third seminar was held in Ottawa in October 2017. Based on these seminar presentations we can get to know the current status of data collection methods followed by various national statistical agencies and the future direction of the data collection methodology.

Australian Bureau of Statistics: ABS has adopted CAPI based survey data collection process for all their surveys. Online forms were a significant part of the data collection strategy for the 2011 Population Census and Housing and the 2011 Agricultural Census. In 2013 online forms were used for 8 business surveys.

Use of administrative data has steadily increased in ABS, taking advantage of introduction of GST. *A single organizational unit has been established to centralize handling of administrative datasets in ABS*. Adoption of technologies like CAPI, spreadsheet based electronic form for offline data collection, and online web data collection has resulted in incremental cost of data collection. The ABS is developing a Metadata Repository and Registry (MRR). ABS is building a single platform for managing interactions with all data providers, whether households, individuals, businesses, or other entities.

Statistics Netherlands: Apart from using CAPI and CATI for survey data collection, Statistics Netherlands (SN) has been using web based survey also. Based on the response rate and response error of web survey, SN has recommended adoption of mixed-mode survey instead of using web only mode of survey. The agency has also created an innovation lab to test new methods, try out non-standard software and simulate alternative statistical process. Some of the research projects initiated by the lab are in the areas of use of Internet as a source of data, use of internet robots for collection of price data etc.

Statistics Canada: In 2010**,** the agency started implementing e-questionnaire as the primary mode of collection for over 160 business and household surveys. It is estimated that e-questionnaire project is to reduce survey cost incurred by the agency by more than $2 million..

CSO , Hungary: The agency launched a web based data collection application called ELECTRA in October 2012.

Federal Statistical Office, Germany: The agency is experimenting with Computer Assisted Web based Interview as a possible mode of data collection.

Office of National Statistics-UK: The agency is trying to use more of administrative data than data collected through survey for creating official statistical outputs. Their approach is- why collect what already exists? The data collection strategy is undergoing a radical transformation. The stated goal is " To rebalance ONS' data collection activity significantly toward wider, more integrated use of administrative data sources, thereby reducing our reliance on large population and business surveys…Further, the remaining survey operations will become more efficient through a move from paper and personal interviews-based collection to online data collection"

US Census Bureau: The agency collected point-of –sales scanner data through a third party from over 1250 retailers representing 300,000 stores and e-commerce platform worldwide. Data such received are compared with data collected by the agency on a regular basis from sampled retailers through Monthly Retail Trade Survey (MRTS) and Annual Retail Trade Survey (ARTS). At the national level the scanner data based estimates were comparable to those obtained through MRTS survey data. The result has shown the great potential of using scanner data for organized retail trade sector for national accounting purpose.

Based on brief summary of data collection strategies / technologies that are unfolding in advanced countries, we may conclude that the Indian Official Statistical system must look into these best international practices and methodologies being used by the statistical agencies of these countries.

Initiatives taken by Indian Statistical Agencies

Although Indian statistical agencies have started online data collection in some areas (ASI, IIP WPI etc.) the current status of data collection is rudimentary to say the least. None of NSSO surveys have been launched in CAPI mode. CSO does not have any system of collecting administrative data electronically from the agencies responsible for capture of data at its lowest level of granularity.

Non-availability of Administrative data of required quality is one of the most important reasons for perceived weakness in the Indian official statistical system. The list of shortcomings of the Indian Official Statistical System that CSO itself has put on its website (accessed on 21st February 2018) is quite revealing and if these are addressed sincerely and purposefully many of the ills of the system would get resolved. The reported shortcomings that fall within the purview of this Committee are noted below:

- Delays in publication of results
- Large and frequent revisions of published results
- Gross discrepancies between official statistics from different sources

- Occasional disagreement between tabulated summary results and publicly available basic data from which the summary has been produced

Finally, the website notes the following:

“Administrative Statistical System has been deteriorating and has now almost collapsed in certain sectors. The deterioration had taken place at its very roots namely, at the very first stage of collection and recording of data, and has been reported so far in four sectors: agriculture, labour, industry and commerce. The foundation on which the entire edifice of Administrative Statistical System was built appears to be crumbling, pulling down the whole system and paralysing a large part of the Indian Statistical System. This indisputably is the major problem facing the Indian Statistical System today” (para no 14.3.10 of http://www.mospi.gov.in/143-administrative-statistical-system)

It is ironical that when all advanced countries are gradually moving towards greater dependence on Administrative Data (as noted above) , Indian official statistical system went on for a “ massive expansion of National Sample Surveys, as a quick means for data collection for GDP estimation” (para 14.3.17 of the CSO website cited above).

Even the surveys conducted by NSSO are still in the PAPI mode. No serious effort has been made to adopt one of the many Computer Assisted technologies that are being used by statistical agencies worldwide. A list of these technologies is given in Annexure

Based on our review of the best international practices and given the rapid progress in adoption of latest Information and Communication Technology this Committee recommends the following measures be taken by the official statistical agencies-both at Centre and the States.

1. CSO/ NAD should get direct access to data captured by departmental statistical agencies or data collecting arm of different regulatory and Developmental bodies

like Tea Board, Controller of Coal et. The method of access would depend on the technological infrastructure prevailing in primary data collecting agency. Some of the possible technologies are : web services, FTP based access, authorized accessible Views of database maintained by the data collecting agency etc. The format, granularity and other metadata of data required by CSO / NAD would be given by them only.

2. To start with CSO/NAD should be able to access raw WPI data and MCA 21 data over a network without any manual intervention
3. CSO should develop a cloud based Data Collecting application managed by it at its data center. This application should have web based architecture and would be made available to all state statistical bureaus to collect data from state agencies engaged in collection of data, a part of which is required by NAD for GDP estimation. The collecting agencies would report only data items required by CSO/NAD and not all data collected by them for administrative use. This online reporting system would ensure quick and validated data flow to NAD much before the same get published in a summary from the publication of the data collecting agency.
4. RBI has already created CSO-National Factsheet on https://dbie.rbi.org.in as per their requirement through which latest data on Trade Statistics (Service Trade), Fiscal Statistics (Fiscal Health Sector), Money & Banking (Banking) data is shared through an API which is having database level access to RBI database of Balance sheets and P&L accounts of banks maintained by RBI. RBI has started using MCA data for its studies on Performance of Private Corporate Sector, the same is also published in RBI data warehouse https://dbie.rbi.org.in. NABARD, IRDA and SEBI / EPFO / PFRDA should also provide database level access to CSO.
5. NSSO should immediately draw a plan for moving all NSSO surveys to a mixed mode data collection methodology to begin with and finally to fully electronic mode of data collection.

6. CSO should work with ERP system vendors like SAP, RAMCO and others to create data in the required format of ASI used in ASI survey. This will significantly reduce time lag in submission of ASI data by large enterprises. CSO should also create a web based ASI data submission process by the factories themselves. Once a Business Register is created as planned, the link between enterprise and establishments can be worked out. This would enable use of automated data integrity verification tools between two sets of data , namely ASI data and MCA21 data.
7. CSO should set up a small group to explore the possibility of using Internet as a source of data. A corpus of fund can be created to award research project to research Institutes like IITs/ISI/IISc to create proof-of concept applications to collect data from Internet.

Need for creation of Statistical Database

A database is a collection of interconnected sets of records where record sets could be files or tables. The main objective of keeping data in a structured database is to ensure consistency and integrity of data so that data can be accessed, managed and retrieved easily. The organization of data in a database is driven by the requirements for which a database is created. For example , an Online Transaction Processing (OLTP) database is different from the corresponding Online Analytical Processing Database (OLAP) although both these databases contain facts about the same objects.

A database should be differentiated from the outputs of the database. Data from a NSSO survey can be stored in a relational base at the lowest level of granularity (i.e. household) but the output tables can be the outputs of some queries addressed to this database. To store these outputs in retrievable files or a database object itself is a technical decision that can be taken depending on the specific requirements.

A statistical database is a database of input data that are used for creating statistical products. Such database is by definition is an analytical database as it only keeps data as a snapshot of past and not as data regarding an object as and when a change occurs in some aspects of that object. There are two ways one can store such data either in the form of a multi-dimensional cube or in the form of normalized relational tables.

It is expected that any online reporting system would store collected data in a proper database. For example, if price data for CPI is collected through a web portal it is expected that a proper database has been created with an underlying data model that would be invariant to change in base year, weighting diagram or basket of commodities. More importantly the related metadata would also be stored in the same database. The robustness of the data model would be tested when we can easily insert new data without changing the underlying data model and retrieve any such data without much effort. For example, we should be able to retrieve all provisional estimates of WPI along with final estimates for each period for which data are available. For example, one should be able to get a report giving variability across various quotations for each commodity for any chosen period.

In summary, there is no statistical advantage by creating an online reporting system with a rudimentary underlying data model.

This Committee recommends that NAD must create a statistical database of socio-economic data it is collecting and using for compilation of GDP

Building a Metadata Repository

Metadata is data about data. It has two main components: (1) Business Metadata, (2) Technical Metadata. The organizing principle for metadata management of a statistical database for a national statistical organization should be the framework provided by SNA.

UNECE has conducted a study of 18 statistical system metadata and the same is available in wiki METIS. Statistics Canada maintains metadata in its Integrated Metadata Database (IMDB).

This Committee recommends that CSO should start without any further delay to create a metadata database and metadata of following three categories should be created:

1. Metadata related to data structure and its contents
2. Process related metadata- including, inter alia, sampling methods, data collection methods, editing process etc
3. Business related metadata- definitions, conceptual hierarchies etc.

Data Dissemination:

Official statistics are a public good. The UN Statistical Commission has laid down 10 fundamental principles of Official Statistics, the first three and the 6th of which are relevant for our discussion. These are reproduced below (emphasis ours).

Principle 1: Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available **on an impartial basis by official statistical agencies to honor citizens' entitlement to public information**.

Principle 2: To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and **presentation of statistical data**.

Principle 3: To facilitate a correct interpretation of the data, the statistical agencies are to present information according to **scientific standards on the sources, methods and procedures of the statistics**.

Principle 6:<u>**Individual data**</u> collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, <u>**are to be strictly confidential and used exclusively for statistical purposes.**</u>

To comply with the above broad principles, dissemination policy of any official statistical system must exhibit the following characteristics:

1. Impartiality: All citizens should have same right of access to official data and no discrimination in favor executive branch of the government should be allowed.
2. Entitlement: Every citizen is entitled to access official statistics subject to the constraint of respect for privacy of legal persons providing the data.
3. Metadata best practices: Complete metadata must be made available to the public
4. Presentation : Presentation of data should be according to the best international practices
5. Privacy and Confidentiality: Identity of data providing citizen must be protected.

The current best practices with regard to data dissemination are illustrated below with examples from 3 leading statistical organizations, namely Statistics Canada, Australian Bureau of Statistics, Office of National Statistics, UK and US Census.

Statistics Canada: The website of Statistics Canada (STATCAN) includes the following:

1. the official release bulletin: The Daily
2. Two large output databases (Canadian Socioeconomic Information Management System (CANSIM), imports and exports)
3. Thousands of data tables
4. census modules, including community profiles, highlight tables, topic-based tabulations, and data visualization tools
5. Major modules for browsing information: "Statistics by subject"; "Information for analysts and researchers"; "Information for survey participants"; "Stay connected" (including blogs, chat sessions with experts, and a video centre

To engage with the users of official statistics STATCAN adopted the following measures:

<u>Website Evaluation Survey</u> (WES): To measure users' satisfaction with the website, STATCAN regularly conducts this survey, identify what has worked and what has

not. Through this survey, STATCAN receives feedback from 10,000 website visitors annually.

Internet Visitor Pattern Analysis (VPA): VPAs use quantitative data (log files) to derive relevant information about visitors' profiles, visit duration, top navigation, content preferences, brand recognition, etc.

STATCAN has taken initiatives to make the website more versatile, interactive, impartial, timely and cost-effective access to information. In February 2012, the CANSIM database and self-standard products became free of charge. License restrictions have been removed as well.

A new user interface for CANSIM launched in February 2012 allows site visitors to have an initial view of the data in fewer than three clicks. From that initial view, the users can customize their tables (layout and information).

Office of National Statistics (ONS)-UK: In 2013, ONS brought out a strategy paper that outlined a roadmap for the organization for the next decade. One of the strategic aims of the organization was to “Dramatically improve the communication of our[ONS’s] statistics and analyses”. The detail plan included the following action plans:

1. develop new ways of presenting our data which enhance our online and digital offering and support greater customer understanding of our statistics
2. open up more of our data for onward use, while ensuring that confidentiality pledges to data suppliers are maintained, and
3. complete the programme to deliver enhanced website capability, open up access to ONS data, and improve our dissemination channels.

The overarching aim of the dissemination strategy was to “achieve the widest possible dissemination and use of official statistics and data by better utilizing the Web and associated technologies and by adopting consistent data and metadata formats and standards”. The ONS decided to have a “single site through which users can gain access to all official statistics”.

Like STATCAN, ONS website allows users to request specific data output according to their needs. Starting from November 2012, the ONS has supplied 2978 user requested reports till 28th November 2017. ONS website also allows users to browse "by subject". For example, Sector Accounts of National Income Accounts can be accessed by two clicks and on this page one can browse time series, dataset, publications and methodology. Under the sub-menu Time series, 4599 datasets are available for viewing and download.

ABS-Australian Bureau of Statistics: ABS has laid down two strategic objectives with regard to dissemination of data, as reproduced below:

1. ABS aims to ensure that data from the same collection for the same reference period and revision, are absolutely consistent whatever the medium of dissemination, also that data disseminated from different collections are relatable to the fullest extent possible and that any inconsistencies are minimal and can be readily identified through inspection of the underlying metadata. An important facility supporting this aim is a corporate repository containing output datasets that can serve Client Services and Subject Matter areas by:

   a) enabling fast and comprehensive searches of output datasets using client oriented terminology
   b) enabling fast retrieval of customized datasets to suit client needs with respect to coverage and data items;

2. The principal objectives of this policy are to ensure that the metadata describing output datasets are fully and consistently defined, that the corresponding data are readily available, and that storage and description of output data are integral to the processes by means of which data from a collection are disseminated.

ABS website provides TableBuilder , an online tool for creating tables, graphs and maps using ABS microdata. The following are the features of this tool :

1. select, customize, create and save tables and customized data
2. display counts, percentages and relative standard errors in your tables
3. calculate means, medians, quantiles and ranges for continuous variables

4. produce output as tables, graphs or maps (maps are currently available for Census TableBuilder datasets only)
5. download tables as CSV, Excel and SDMX files
6. download graphs or maps as PDF and KMZ files
7. create, save and share customized geographic areas and recodes with other users (registered users only).

ABS website also provides access to Confidentialised Unit Record Files (CURFs) - microdata files containing individual responses, which have been treated to protect privacy and confidentiality using a range of statistical techniques. Access to CURF is enabled through secure online data query service called Remote Access Data Laboratory (RADL). There is another facility called DataLab that allows power users to undertake interactive complex analysis of microdata.

US Census Bureau: US Census Bureau website includes American Factfinder, an advanced search tool that can search all available census data using native functionality of the tool. Search of data tables can be across categorical attributes of data tables like Topics, Geographies, Race and Ethnic Groups, Industry Codes and occupation Codes.

In summary, all national statistical organizations are providing interactive access to database of their statistical products and allowing user defined report creation facilities. This is possible only when the metadata of all statistical products are in a queryable database and data elements are also stored in such a database. As more and more analysts and researchers are becoming comfortable with online access to and viewing of data, creation of such a user friendly interactive access to data is a must for any responsive and user-oriented statistical organization.

This Committee recommends the following with regard to adoption of best practices in data dissemination by CSO.

1. Dissemination of statistics by CSO should be in a machine readable and computer processing capable format. Dissemination in pdf format alone is not acceptable.

2. Any disseminated statistical product must be hyperlinked to its corresponding metadata.
3. A queryable catalogue of statistical products should be made available.
4. At any pointy of time a user should be able to get information data available on a particular topic.
5. It should be possible to navigate to another Indian statistical agency’s data disseminating website when a user submits a relevant query.

Summary and Conclusions

To ensure timeliness and trust in official statistical outputs, data need to be collected and recorded electronically at the first stage of data collection. This requirement is valid for both administrative data as well as survey data. Although electronic mode of data collection is gradually being adopted by the Indian official statistical system both at the centre at state level, level of automation at all stages of lifecycle of official data is yet to reach a satisfactory level. One of the major shortcomings of the prevailing data management practice of the Indian OSS is the absence of any automated data flow mechanism between various statistical agencies. This problem is compounded with the absence of database oriented data storage system for many important datasets. Even when data are stored in a database that supports standard SQL queries, lack of proper metadata is a serious handicap for policy makers and end users in accessing data required by them.

To address the above noted weaknesses in the Indian OSS, the committee has made a number of recommendations listed below. The committee believes that a mission mode approach should be taken to bridge these gaps within a reasonable time line.

List of Recommendations

1. Data required by CSO/NAD, for compilation of national accounts, from other statistical agencies including central ministries, departments and state statistical bureaus should be made electronically available to NAD without any time lag. NAD can prepare a specific return format for such data submission by the respective agencies and obtain it over a secure network. The data should be validated at the time of data submission itself and stored in a proper statistical database
2. The data should be validated at the time of data submission itself and stored in a proper statistical database

3. Wherever possible, CSO/ NAD should get direct access to data captured by departmental statistical agencies or data collecting arm of different regulatory and Developmental bodies like Tea Board, Controller of Coal et. The method of access would depend on the technological infrastructure prevailing in primary data collecting agency.
4. To start with CSO/NAD should be able to access raw WPI data and MCA21 data over a network without any manual intervention
5. CSO should develop a cloud based Data Collecting application managed by it at its data center. This application should have web based architecture and would be made available to all state statistical bureaus to collect data from state agencies engaged in collection of data, a part of which is required by NAD for GDP estimation. The collecting agencies would report only data items required by CSO/NAD and not all data collected by them for administrative use. This online reporting system would ensure quick and validated data flow to NAD much before the same get published in a summary from the publication of the data collecting agency.
6. NABARD, IRDA and SEBI / EPFO / PFRDA should also provide database level access to CSO.
7. NSSO should immediately draw a plan for moving all NSSO surveys to a mixed mode data collection methodology to begin with and finally to fully electronic mode of data collection.
8. CSO should work with ERP system vendors like SAP, RAMCO and others to create data in the required format of ASI used in ASI survey. This will significantly reduce time lag in submission of ASI data by large enterprises. CSO should also create a web based ASI data submission process by the factories themselves. Once a Business Register is created as planned, the link between enterprise and establishments can be worked out. This would enable use of automated data integrity verification tools between two sets of data, namely ASI data and MCA21 data.

9. CSO should set up a small group to explore the possibility of using Internet as a source of data. A corpus of fund can be created to award research project to research Institutes like IITs/ISI/IISc to create proof-of concept applications to collect data from Internet.
10. This Committee recommends that CSO should start without any further delay to create a metadata database and metadata of following three categories should be created:
    1. Metadata related to data structure and its contents
    2. Process related metadata- including, inter alia, sampling methods, data collection methods, editing process etc
    3. Business related metadata- definitions, conceptual hierarchies etc
11. This Committee recommends that NAD must create a statistical database of socio-economic data it is collecting and using for compilation of GDP. It is emphasized that there is no statistical advantage in creating an ORS without a proper underlying data model
12. Dissemination of statistics by CSO should be in a machine readable and computer processing capable format. Dissemination in pdf format alone is not acceptable.
13. Any disseminated statistical product must be hyperlinked to its corresponding metadata.
14. A queryable catalogue of statistical products should be made available.
15. At any pointy of time a user should be able to get information data available on a particular topic.
16. It should be possible to navigate to another Indian statistical agency’s data disseminating website when a user submits a relevant query.

Annexure I

Methodologies for Collection of Survey Data

Survey Data

The traditional method of collection by conducting personal interviews or telephonic survey followed by recording of them in paper based forms is gradually being replaced. The process of replacement is being driven both by technology and the preferences of respondents. The various methodologies for conducting a survey can be classified across following dimensions of a survey

1. Administrator of a survey questionnaire
    a. Interviewer
    b. Self-administered
2. Data capture method
    a. Paper based
    b. Electronic (on a device like tablet, laptop or mobile)
    c. Mixed (paper along with audio /image captured on an electronic device)
3. Medium
    a. Face to face
    b. Telephonic interview by a human interviewer
    c. Mail
    d. IVR
    e. Web
4. Visual /audio dimension
    a. Visual
    b. Audio
    c. Mixed

Based on these dimensions, a possible classification of existing survey methodologies is given in Table 1 below.

Table 1: Classification of Survey Methods

Survey Methodology	**Mode of administration**	**Data Capture method**	**Medium**	**Visual/Audio**	**Acronym**
Face to Face-traditional method: Pen and paper interview	Interviewer	Paper	Face to face	Audio	PAPI
Face to Face with electronic device	Interviewer	Electronic (Tablet/Laptop /Mobile)	Face-to-face	Audio	CAPI
Computer assisted self-interview	Self-administered	Electronic	IVR/Web	Audio / Visual	CASI
Telephonic interview	Interviewer	Paper	Telephone	Audio	TI
Computer Assisted Telephonic interview	Interviewer	Electronic	Telephone	Audio	CATI
Mail	Self-administered	Paper	Mail	Visual	Mailed Survey
Telephone audio computer-assisted self-interview	Self-administered	Electronic	Telephone	Audio	T-ACASI
Online Questionnaire	Self-administered	Electronic	Web	Visual	OQ
E mailed Questionnaire	Self-administered	Electronic	Web	Visual	EQ
Computer Audio Recorded Interview	Interviewer	Electronic	Face-to-face	Audio	CARI
Web form	Self-administered	Electronic	Web	Visual	WSAQ

To the above alphabet soup of survey methodologies, we may add another important dimension of survey; that is the underlying survey design. Traditionally, respondents or the Final Stage Units have been selected through a probabilistic design that makes the collected data amenable to estimation of target population parameter(s) with desirable properties like unbiasedness , consistency etc. But with the currently available technology it is possible to administer a survey to a very large segment of the population without any proper sampling design. Pollsters like Gallup have used random digit dialing (RDD) technique to conduct surveys. The growing popularity of algorithm oriented data analysis can supplement survey design deficiency.

For this committee's purpose, by "Online Reporting System" we would refer to those methodologies that capture data electronically at the time of administration of a survey. Given the current status of infrastructure of the country and high incidence of digital poverty, official statistical system needs to carefully calibrate transition to digital mode of data capture in a phased manner. In the Indian context we should focus our attention only two modes to begin with, namely CAPI and CARI. These two modes can be thought as complementary to each other and can be deployed simultaneously for any survey.

Computer Assisted Personal Interview (CAPI): CAPI was developed with the same four objectives as its predecessor CATI was:

1. Increased efficiency in terms of time taken to collect, collate and process survey data
2. Improvement in survey data quality
3. Reduction in cost per unit of data items collected and processed
4. Better and more effective interaction between respondent and interviewer while administering complex questionnaire

The first difference between the traditional PAPI and CAPI lies in the replacement of Paper with Computer for data capture. In computer-assisted personal interviewing (CAPI) the interviewer reads preloaded questions from the laptop/tablet / mobile screen to the respondent. The respondent's answers are immediately entered into the device, thereby eliminating any further data entry. Given today's technology it is possible to automate routing through the questionnaire based on answers given. In-built consistency checks can generate automatic real time alert to the interviewer, so that anomalies can be resolved with the respondent. The envisaged benefit of CAPI over PAPI has been extensively examined in the context of developed countries. Initial studies did not find any major resistance to introduction of CAPI by the survey investigators. After analysis of interviewer attitude to introduction of CAPI, one early study in respect of US interviewer for Current Population Survey noted the following:

Interviewer attitudes appear to become more positive with experience using CAPI. Interviewers' initial experience with the new technology does not appear to dampen their enthusiasm. This suggests that many who are initially skeptical may be won over once their early fears or concerns about CAPI are allayed (Couper and Burt 1994).

For a developing country, one study reports the relative merits and demerits of CAPI over PAPI based on a randomized experiment of 1840 households on the Island of Pemba in Tanzania. The paper finds that that errors leading to missing variables in PAPI are virtually eliminated in CAPI. Most interestingly the study observed that "paper questionnaires can lead to estimates of higher mean consumption, lower poverty and higher inequality". On the other hand automatic consistency checks may motivate to enter any suitable data to get around the alert message without verifying the same with the respondent. Only stringent quality checks can dis- incentivize such actions on the part of interviewers.

Adoption of CAPI: The Central Bureau of Statistics of Netherland was the pioneer in adopting CAPI technology for its regular surveys. Since then many statistical agencies have adopted either CAPI or CATI for conducting its regular survey. In fact, the statistical agencies in developed countries are gradually moving towards online data collection over internet

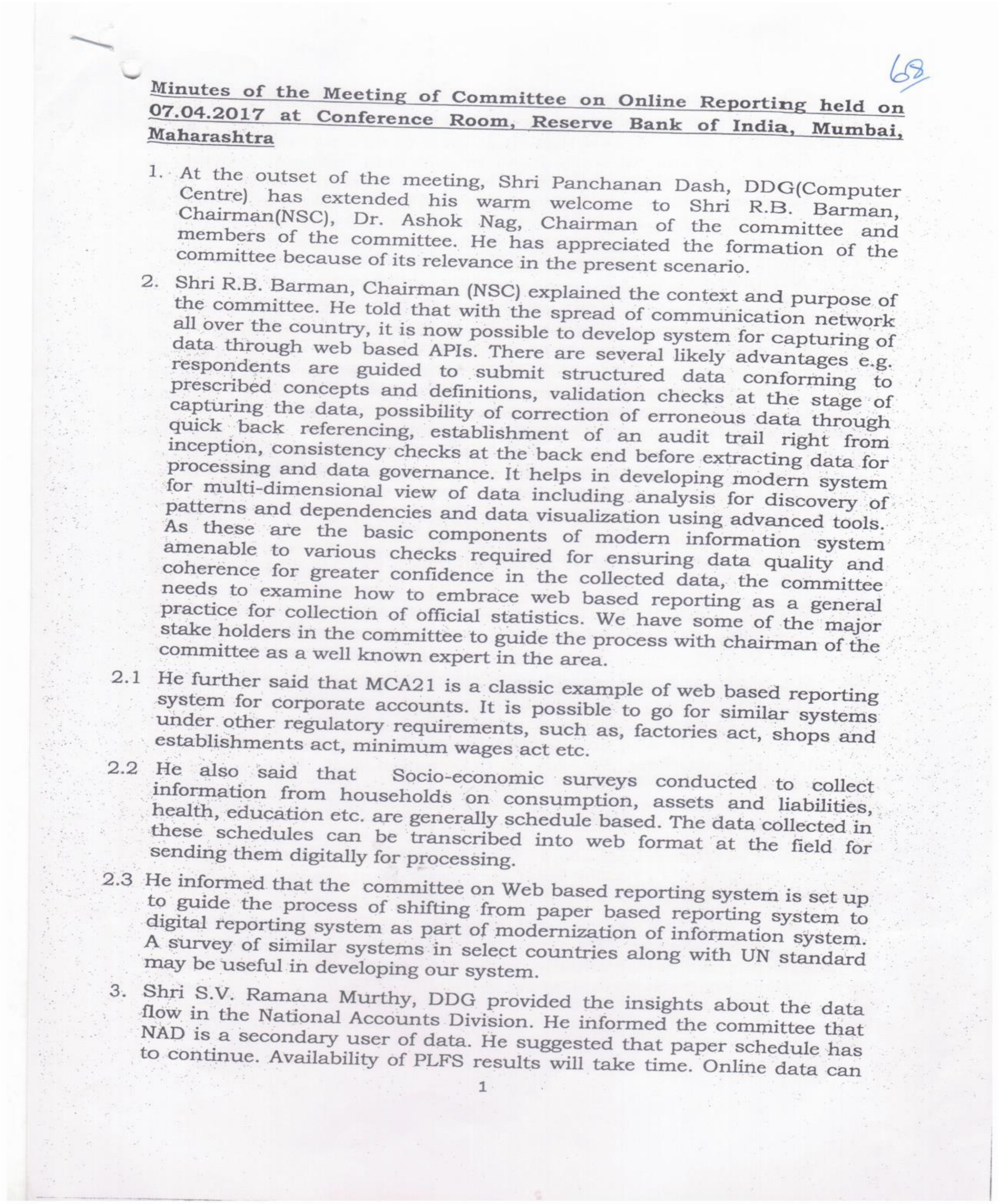
Computer Audio Recorded Interview (CARI): CARI should be looked upon as an additional facility for CAPI but it could be also combined with PAPI also. The main components of a CARI system are:

1. Full audio / screen images (in case of CAPI) recording of survey,
2. Transmission of the audio/ image file to a central server for monitoring purpose
3. For quality control offline review of transmitted files to detect communication failure between interviewer and respondent leading to erroneous response
4. Feedback to interviewer and survey designers for quality improvement

It is now possible to deploy advanced voice recognition technology to capture data independently and compare with the one entered by interviewer. A small portable tablet would be sufficient for this purpose.

ANNEXURE 2

Minutes of the Meeting of the Committee held on 07.04.2017 at Mumbai



68

Minutes of the Meeting of Committee on Online Reporting held on 07.04.2017 at Conference Room, Reserve Bank of India, Mumbai, Maharashtra

1. At the outset of the meeting, Shri Panchanan Dash, DDG(Computer Centre) has extended his warm welcome to Shri R.B. Barman, Chairman(NSC), Dr. Ashok Nag, Chairman of the committee and members of the committee. He has appreciated the formation of the committee because of its relevance in the present scenario.

2. Shri R.B. Barman, Chairman (NSC) explained the context and purpose of the committee. He told that with the spread of communication network all over the country, it is now possible to develop system for capturing of data through web based APIs. There are several likely advantages e.g. respondents are guided to submit structured data conforming to prescribed concepts and definitions, validation checks at the stage of capturing the data, possibility of correction of erroneous data through quick back referencing, establishment of an audit trail right from inception, consistency checks at the back end before extracting data for processing and data governance. It helps in developing modern system for multi-dimensional view of data including analysis for discovery of patterns and dependencies and data visualization using advanced tools. As these are the basic components of modern information system amenable to various checks required for ensuring data quality and coherence for greater confidence in the collected data, the committee needs to examine how to embrace web based reporting as a general practice for collection of official statistics. We have some of the major stake holders in the committee to guide the process with chairman of the committee as a well known expert in the area.

2.1 He further said that MCA21 is a classic example of web based reporting system for corporate accounts. It is possible to go for similar systems under other regulatory requirements, such as, factories act, shops and establishments act, minimum wages act etc.

2.2 He also said that Socio-economic surveys conducted to collect information from households on consumption, assets and liabilities, health, education etc. are generally schedule based. The data collected in these schedules can be transcribed into web format at the field for sending them digitally for processing.

2.3 He informed that the committee on Web based reporting system is set up to guide the process of shifting from paper based reporting system to digital reporting system as part of modernization of information system. A survey of similar systems in select countries along with UN standard may be useful in developing our system.

3. Shri S.V. Ramana Murthy, DDG provided the insights about the data flow in the National Accounts Division. He informed the committee that NAD is a secondary user of data. He suggested that paper schedule has to continue. Availability of PLFS results will take time. Online data can

1

ANNEXURE 3

Format: Data Source Assessment for Online Reporting System

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
	Data Source Summary		
1.	Name of Data Source/ Data System/MIS	Mandatory	MCA 21 System
2.	Name of Agency/Department collecting/custodian of Data	Mandatory	Ministry of Corporate Affairs (MCA), Government of India
3.	Name of Agency /Department owning Data	Mandatory	Ministry of Corporate Affairs (MCA), Government of India
4.	Purpose and use for collecting Agency /Department	Mandatory	Monitoring and Regulation of companies registered under Companies Act.
5.	Name and address of contact person in Agency /Department (with contact No and email)	Mandatory	Amardeep Singh Bhatia, Joint Secretary (eGovernance), Room No. 505, A WingShastri Bhawan, M/o Corporate Affairs, New Delhi-110001

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
6.	Description of Data source	Mandatory	Few Illustrative information and their respective e-forms are: • Incorporation Form (Form 1) • Annual Return (as on AGM date) - Share Holding (20B) • Financial Statements (as on B/S date) (23AC, 23ACA) ○ Subsidiary structure ○ XBRL for selected ○ Auditors Comments • Director's Details (DIN 3) • Director's Association (Form 32)
7.	Key Data Fields	Mandatory	As per Annexure-1 Schema
8.	Technology Stack (Database system/ Data warehouse)		SAP CRM, SAP BI, OpenText IBM WebSphere, IBM DB2, SQL Server (Document Metadata)
9.	Number of records collected annually		More than 9lakh active companies
10.	Data size (in MB) collected annually (both Structured & Unstructured)	Mandatory	700 MB (Structured) 4000 MB (Unstructured Documents)
11.	Date of commencement of collection		Sep 2006

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
12.	Cumulative Data size (in MB) collected	Mandatory	3500 MB (Structured) 32000 MB (Unstructured Documents)
	Data Collection Process		
13.	Is there a legal framework for collection of data? If yes, please give details	Mandatory	Yes, Under Section 610B of the Companies Act, 1956 and Rules thereunder.
14.	Is there a penalty/fine for late or non-submission of data? If yes, please give details of legal provisions and amount of fine imposed		Yes, penalty for non-submission of information, etc. are prescribed in relevant sections of the Companies Act, 1956 such as under sections 162, 614, 629A, etc.
15.	Timing and frequency of data collection (Monthly/Quarterly/Yearly)	Mandatory	As prescribed in the relevant sections of the Act.
16.	Is the list of information submitters/ Directory of reporting entity available?	Mandatory	Yes
17	Is there a defined process for data collection? If yes, please give details of process flow from field formations/hierarchy (Add Annexure if needed)		Yes, companies file the information through designated e-forms using DSC based authentication of Authorized Signatory of the Company.

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
18	Whether data structure/ schema is prescribed? If yes, please give details	Mandatory	Yes, as per the prescribed e-forms. Details as per Annexure-1.
19	Whether data validation rules are prescribed at data submission stage? If yes, please give details		Yes, required validation checks are built-in the e-forms.
20	Whether data preparation utility is provided to the reporting entity? If yes, please give details		No, companies have to fill-in the designated e-form and submit to MCA21 through Front Office facilities.
21	Is there a defined process for data validation? If yes, please give details (both Online & Offline)		Yes
22	Whether data validation results are shared with reporting entity? If yes, please give details		For XBRL filings, MCA has provided a Validation Tool that identifies inconsistencies and inform the filer to correct and resubmit.
23	Is there a defined process for pursuing non-filers or late filers? If yes, please give number of cases in last 2 years.	Mandatory	Yes, as per provisions of the Companies Act, 1956.
24	Whether any quality assurance procedures are performed? If yes, please give details whether it is Electronic or Manual		Yes, back-end validation about company details, signatory details, etc. are done for each e-form.

S. No.	**Question**	**Mandatory/Optional**	**Illustration (M/o Corporate Affairs)**
25	Is a data dictionary/codebook (such as Instructions to reporting entity) available? If yes, please give details		Taxonomies for XBRL filings, and designated e-forms for other filings.
26	Are any modifications planned in any aspect of data collection, processing, etc.		New forms as per provisions of the new Companies Act, 2013. Implementation of MCA 21 version 2 using SAP CRM.
27	Whether an HelpDesk is available to reporting entity for resolving technical/functional/ operational grievances		Yes, Corporate Seva Kendras are available for filers regarding any issue in submission of forms. Online reporting & tracking of issues is also available.
	Usefulness of Data Source		
28	How has data been used by Agency/Department owning data?		Monitoring Non-filers, Identifying non-compliance, Monitoring & Administration of field offices, macro-economic policy making inputs, regulatory analysis by MCA officials, etc.

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
29	What are some opportunities for using the data in other departments?		MCA21 data provides a 360 degrees view on working and status of all companies registered in India. Specifically, all relevant information such as list of Active companies, their Financial Information, their Holding subsidiary relationship, Director Company association, their Shareholding Pattern, various compliances, etc. are available in MCA21 system.
30	What are the main limitations of the data?		Information furnished as scanned/pdf documents are difficult for automatic analysis by the machine.
	Data Access and Exchange		
31	Are there written laws/rules/policies and/or procedures on sharing data with other Departments? Provide details of Data sharing Policy (NDSAP), Information Technology Act 2000 and Aadhar Act 2016.	Mandatory	All public documents pertaining to a company are available for viewing/sharing as per provisions under section 610 of the Companies Act 1956. The facility of View Public Document has been enabled in MCA21 system to implement this provision of the Act.
32	What are the confidentiality requirements related to data sharing with other Departments?		Modalities are being worked out.

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
33	Whether a unique ID is maintained for each individual or entity? If yes, please give details. How is de-duplication done?	Mandatory	Yes. Company Identification Number (CIN) for companies and Director Identification Number (DIN) for directors.
34	What information is needed to identify (search criteria) a specific person or entity?		For Director (Name, DoB, DIN, PAN), and for Company (Name, CIN)
35	What information is required in a request for specific information?		Modalities need to be defined. However, any data request should have Unique identifiers, Period, Person making request, Reason for Request, Priority, etc.
36	Is there a standard template for receiving request for information?	Mandatory	Needs to be developed.

S. No.	Question	Mandatory/Optional	Illustration (M/o Corporate Affairs)
37	What types of data can be suo-moto shared with other Department spontaneously? Whether Web Services/API/ Secured File Transfer Protocols (SFTP), etc. have been developed by the Department.	Mandatory	MCA and CBDT may share information relating to company/director that may assist in identification of non-compliance of provisions of companies laws/tax laws, and thereby improve regulation and decision-making. Few examples of information sharing are of cases having tax implications (above predefined threshold) such asFalsification of accounts, Siphoning of funds, Share capital related, Unsecured loan, Bad debts, Directors remuneration, etc.
38	What is the preferable mode and frequency of sharing bulk data with other Department?		To be discussed.
39	Are you aware about the provisions regarding data collection in the Collection of Statistics Act 2008, and Rules thereof.	Mandatory	

References