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.

1. Introduction

Indian Official Statistical System (OSS) is organized within the broad administrative and political framework provided in the Indian Constitution. The federal structure of Indian constitution provides the jurisdictional areas of administrative power to be exercised by the Government of India and the State Governments, either exclusively or jointly. The authority for collection of statistics on any subject area is guided by the status of the subject area in the Union, State and Concurrent list. The following features of the Indian Statistical System that the Rangarajan Commission had noted in their report published in 2001 are still relevant and valid.

(i) The Administrative data is the main source of data
(ii) It is laterally and vertically decentralized
(iii) For most of the sectors, the states are primarily responsible for all activities related to management of official statistics including, inter alia, collection, compilation and report generation.
(iv) The all-India level statistics are mostly aggregates of state level statistics.

It is apparent from the above observations that the Indian OSS is divided into silos, each with its own information architecture and lacking interoperability across the entire spectrum of datasets. Although, administrative measures were taken to consolidate all official statistics under the supervision of central statistical offices at the Centre and at the states, by creating CSO at the Centre and State Directorates of Economics and Statistics in the States, decentralized and departmental nature of data repositories continue to afflict the Indian OSS.

Even when data is collected on a single subject with a regular frequency, users do not get access to a proper database that facilitates creation of user defined analytical report downloadable in standard format. Annual Survey of Industries (ASI) data, for example, is a rich source of data on inputs and outputs of organized manufacturing sector.CSO obtains them in machine readable format. It should not be difficult to create a Data Integration Layer incorporating data quality checking process to ensure that validated data gets loaded into a database that would facilitate data

presentation and analytical reporting. This would significantly reduce “time to release” of data for policy makers and general public. This observation apply to many other important datasets like Index of Industrial Production data, Consumer Expenditure Survey data of NSSO, Budget data of the Central and State governments and so on.

Another major deficiency of all publicly available datasets is that none of them provide metadata to end users as and when data is consumed. For example, NAS is based on a detailed methodological document that is available on CSO website as a downloadable PDF document. Today’s technology allows providing of relevant metadata associated with any data as and when users are consuming that data.

The National Statistical Commission (NSC) decided to constitute five professional committees to assist it on various technical issues as above that pertain to OSS (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. Specifically, these committees should look into data governance issues relating to quality, timeliness and credibility of collected data and derived estimates and make suitable recommendations. NSC expects that these committees should explore the feasibilities of making data and estimates available at disaggregated level by various dimensions of industry, geography, time, size, class etc., and make recommendations towards their realization.

(i) Committee on Real Sector Statistics
(ii) Committee on Financial Sector Statistics
(iii) Committee on Fiscal Statistics
(iv) Committee on Online Reporting System
(v) Committee on Analytics

The committee on Analytics was constituted with Prof. N. L. Sarda, Emeritus Fellow, IIT Bombay, as its Chair. The list of members is given at the end of this report separately

The terms and references of the committee are as under:

(i) To survey best practices on repository for National Statistical System

(ii) To review existing system of data collection, collation and dissemination in Central Statistical Office, National Sample Survey Office and Directorate General of Commercial intelligence & Statistics from the point of view of Analytics. At least one state may also be considered, if feasible.

(iii) To recommend suitable measures to strengthen systems and processes for (1) Data governance and (2) Multidimensional view of data for core statistics forming parts of National Accounts, specially Corporate Statistics, Fiscal Statistics, Agriculture Statistics, etc;

(v) To suggest broad approach to Technology Architecture for data repository on Core Statistics flowing into CSO, NSSO and DGCIS and Application Software and Hardware Procurement Process;

(v) To guide and supervise a Pilot Project to be undertaken by the above concerned departments for hands on experience for building analytics by these departments. Possibility of building a full-fledged repository out of the pilot may also be considered

In order to help the committees to adopt a more focused approach in their deliberations and recommendations, a meeting of the Chairmen/ Co-chairmen/ Member Secretaries of the five committees was held with Deputy Chairman of NITI, Aayog on 26th October, 2016 at NITI, Aayog. The main agenda of the meeting was to have an interaction and explain to them the major objective of an integrated National Statistical System through a bottom up approach and the responsibility of each committee as set out in the terms of reference to help achieve the objective. Dr. R B Barman, Chairman, National Statistical Commission, pointed out that the committee on Analytics is uniquely placed as its recommendations would impact the end users of all statistical outputs of the OSS and thus would determine the success of the modernization effort of the Indian OSS that NSC is striving for.

The committee while deliberating on its mandate felt that the concept of Analytics needs to be properly articulated in regard to official statistics so as to keep the committee's work and recommendations focussed. As data is at the core of statistics, the committee felt that concept and application of analytics must be related to the entire lifecycle of data involved in production of official statistical outputs. The following articulation of the concept of Analytics was accepted as the guiding principle of the committee's work.

Analytics covers the complete life cycle of data, right from collecting to cleansing, organizing, storing, analysing and governing the data. It also includes tools and techniques for analysis of the data.

2. Nature of data and sources of data

India is well known for its richness of official statistics collected in a systematic and scientific way on almost every aspects of its more than billion plus population. India has adopted SNA framework for its national accounting statistics since independence. All important economic indices and national accounts are regularly revised to take into account the changing structure of production, consumption pattern and technology environment. However, such a rich cornucopia of statistical outputs is conducted mostly in silos and requires an integrating framework to make them relatable and amenable to analytical processing of data. In way of illustration, details of some major statistical outputs are given below.

Price Indices

Name of the price index	Producer Department	Market Segment covered	Commodity Basket	Current Series Base Year	Geography Granularity	Frequency	Data Availability /Format
Wholesale Price Index (WPI)	Office of the Economic Adviser Ministry of Commerce and Industry-GOI	Wholesale market-Ex-factory sale Agriculture mandi price, Ex-mines sale	Primary articles, Fuel and Power Manufactured product;	2011-12	All-India	Monthly	Commodity level/ EXCEL
Consumer Price Index (CPI)+ CPI for Rural and CPI for Urban	CSO; Ministry of Statistics and Programme Implementation	Retail market- household purchase for consumption	Goods and services consumed by households with 3 level hierarchies-Group, sub-group sections	2012 (calendar year)	All India/ Rural/Urban/ State	Monthly	State/ Commodity Group and sub-group level/ EXCEL
Consumer Price Index for Industrial Workers	Labour Bureau Ministry of Labour and employment-GOI	Retail market in 78 selected industrial /plantations/mining/ centres spread across India	Retail items consumed by workers	2001	Centre wise / aggregate to all India	Monthly	Group / sub-group / Centre / All India level EXCEL
Consumer Price Index for Agricultural Labour + CPI for Rural Labour	-do-	Rural Retail markets Prices used are the same for two indices but weighting diagrams differ	Consumption items of agricultural rural labourers with 2 level hierarchies – group and sub-group	July 1986 to June 1987	State	Monthly	State wise /month wise/group / sub-group EXCEL

The following observations are made about current information repository structure of different Price Indices:

1. All data are managed and disseminated by respective producing departments.
2. Some states also have their own system of collection of price data and bring out their own state specific price indices. There is no uniform methodology that these states follow, resulting in lack of definitional consistencyand harmony in these series. As these data are required for compilation of state and national GDP, it is important to bring uniformity in compilation methodology of these data series and create acentral repository to store these data in a structured database.

National Accounts:

The CSO publishes National Accounts Statistics (NAS), its flagship statistical product, on annual basis. Yearly tabular reports have been published and these are downloadable either in EXCEL or PDF formats. The NAS has so far been revised with a time gap of a decade, which is being now proposed to be done in every five years., starting with 2011-12 series. Revisions are made in terms of enhanced coverage of current data, improvement in methodological framework and making it consistent with revised SNA and base year changes for constant price series. Metadata about national accounts are available in separate documents. With the release of every new series, CSO brings out a compatible series of important macroeconomic aggregates like GDP, Capital Formation etc. after a time gap. From an end user perspective, analytical utility of data is limited as each national accounts report is published as a stand-alone report and not directly relatable to each other

Annual Survey of Industries:

CSO conducts Survey of Industries annually and the results are published in two volumes. The volume 1 is published in PDF format and 8 major tables included in Volume 1 are also published individually in PDF formats. The users can select the reference year and get the corresponding tables.

The volume 1 provides data on economic parameters pertaining to industries in tabular formats. The volume 1 includes 12 tables. The volume II provides detailed data on materials consumed and value of products and by-products at all-India and at the State levels. The volume II is available only CD-ROM media.

In the absence of a proper database storage that facilitates creation of analytical report downloadable in standard analysis format, utility of such data becomes limited. Given that currently ASI data flows to CSO in machine readable format, it should be eminently possible to create a Data Integration Layer incorporating ETL process and data quality checking process to automate populating the main database that would facilitate data presentation and analytical reporting. This would significantly reduce "time to release" of data for policy makers and general public.

Corporate Sector Statistics:

Historically, the Reserve Bank of India (RBI) was the main compiler of corporate sector financial Statistics based on data made available in Balance Sheet and P&L Accounts of these companies. Although RBI followed a consistent methodology in compilation of these data, its coverage was found inadequate for generating estimates for the entire sector. With the introduction of MCA21 the Ministry of Corporate Affairs started receiving data in XBRL format from all companies registered under the Companies Act. RBI has also 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. As a result the coverage of reported data has increased significantly.

One important shortcoming of the current compilation method is that there is no integration of establishment based data of ASI and enterprise level data of MCA21. Such integration would give much better insight of functioning of corporate sector within the consistent framework of national income accounting. This would significantly enhance analytical use of both the datasets.

Agricultural Sector Statistics:

Agricultural sector is the most important sector for livelihood of the majority of Indian population. For compilation of GDP from the sector, data are needed on many items including but not limited to land use statistics, crop yield, cost of cultivation , farm gate prices . Most of these data are collected at state level and only summary information flow to the central ministry. There is no integrated framework that can tie together these state level data as distributed data with identical data schema. Such integration will empower policy makers and researcher to quickly identify any incipient stress in the sector or emerging potential of any region/crop, apart from many other important policy analyses.

Other Datasets

Many of the above observations apply to many other important datasets like Index of Industrial Production data, Consumer Expenditure Survey data of NSSO, Budget data of the Central and State governments and so on.

One major deficiency of all above datasets made available to data consumers is that none of them provide metadata to end users as and when data is consumed. Give today's technology it should be possible to provide such definitional details as and when the data is consumed.

3. International Practices

The complexity of Official Statistics is significantly more as compared to enterprise level data, even if the enterprise is as large as Exxon Mobil or as diverse as GE. The complexity arises due to the number of subject areas to be covered and the wide variety of facts and attributes thereof to be captured. Even for a small country like Finland the effort required for integrating official statistics is substantial. National Statistical Institutes (NSI- counterpart of CSO) of many countries are now seized with this effort to create maximum value of official statistics which is considered as public good. A few of these exercises are briefly described below.

Statistics Spain

The modernization effort that Statistics Spain launched few years back started with the fundamental realization that “Official Statistics production needs a unified combination of statistics and computer science”. In other words without creating a very robust technological infrastructure the mass of official statistics cannot be utilized to its fullest potential. Realizing this organization has first created a modern system of metadata, in particular of process metadata. For a given process involved in production of statistics, the process is described according to their input(s), output(s), throughput(s), documentation, tools and responsible unit(s). The computer science design principles that have been used to accomplish this are functional modularity, hierarchy and layering.

Australia

Australian Bureaus of Statistics (ABS) started a Statistical Business Transformation (SBT) program in 2015. One of the key objectives of the SBT program is to have increased data integration, bringing together existing data sources or collections efficiently and safely. ABS is in the process of developing a corporate Metadata registry and Repository to centrally store and easily access all statistical metadata. This central registry would help develop an end-to-end metadata driven a Statistical Workflow Management System

Finland

About 95 per cent of all the data that are used in statistical production at Statistics Finland come from administrative data or registers. Statistics Finland, therefore, started building a centralized system for administrative data collection. To achieve this, Statistics Finland ensured that data transfer from administrative data source to Statistics Finland’s repository is controlled by process metadata and transferred data flows to the right directory automatically. Both “push” and “pull” methods are used to manage data transfer.

4. A few case studies from India

Gujarat Integrated Statistical System- GISS:

"Gujarat Integrated Statistical System- GISS" is an initiative by Government of Gujarat to develop a central data repository of the state. The GISS is an integrated data base of all sectors along with an access to high end analytical software. It is accessible to all Government Departments and officials at the state, district, and sub district level. The Gujarat Government has also a develop another portal "Village Profile and taluka planning atlas", a GIS based decision support system developed and institutionalized for spatial planning process in the state through an acceptable , adoptable and affordable GIS based decision support system and visualization at various level of hierarchies.

Maharashtra State Data Bank:

This is an IT initiative taken up by the State Government to consolidate and collate data sets available with the various state departments. The main objective of this initiative is to create a decision support system which will also serve as a knowledge repository for various information seekers such as researchers, academicians and general public.

Open Government Data (OGD) Platform India (http://data.gov.in):

The Government of India (GOI) has formulated the National Data Sharing and Accessibility Policy (NDSAP). The Ministry of Electronics & Information Technology (MeitY) is the nodal Ministry to implement the policy.MeitY through NIC has set up the Open Government Data (OGD) Platform India (https://data.gov.in/) to provide open access by proactive release of the data available with various ministries/ departments/ organizations of Government of India.OGD Platform can be used by government departments to publish their datasets through a predefined workflow. The platform provides APIs through which direct and dynamic query can be made to access data items of selected datasets. OGD Platform also has a Communities component which facilitates forming of communities around datasets.

National Data Registry for NSDI

Department of Science & Technology is in the process of developing a National Data Registry (NDR) forNational Spatial Data Infrastructure (NSDI) to establish and maintain data and service catalogues. NDR will provide standardized catalogue services capable ofgetting consumed by end users or applications.The building blocks of the proposed central repository called Registry would be Registers which are files containing identifiers assigned to items with descriptions of theassociated items. These files contain feature data sets / images and other textual and numerical spatial data. A metadata driven NDR would allow fully automated search, querying, and processing.

Database of Indian Economy (DBIE)

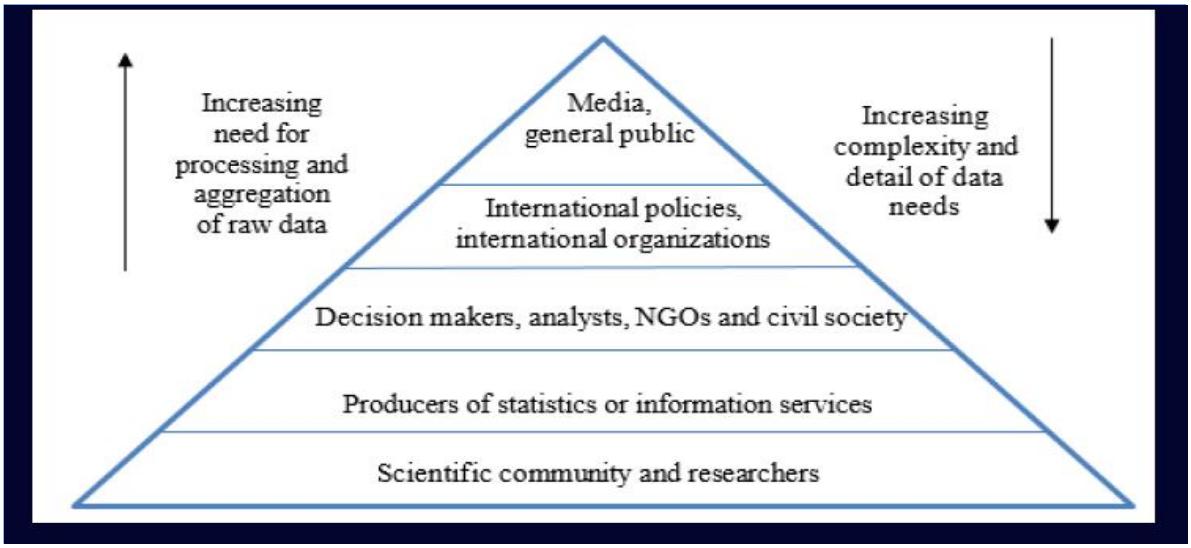
RBI launched an Enterprise Data Warehouse initiative titled Central Database Management System (CDBMS) in early 2000 and released it to RBI's internal users in the second half of the year 2002. This is one of the earliest initiatives of any central bank to establish a Data warehouse of all data available in various departments of a central bank. Based on this DW, RBI released a large portion of data for public through a portal titled DBIE. The broad architecture of CDBMS is given in the Annexure 2.

5. *A Broad Approach to Build a National Integrated Data System (NIDS)*

In April 2014, the Conference of European Statisticians (CES) held a seminar on "What is the value of official statistics and how do we communicate that value?" In that seminar Prof. David J Hand made a presentation in which he very succinctly described the pyramid of stakeholders and their expected requirements from the official statistical system. It is highly pertinent to reproduce that pyramid in this report to be submitted to the National Statistical Commission as the principal motivation to create these committees is to maximize the social return on the investment made in manpower and resources to produce such

statistics. Dr. R.B.Barman, the current chairman of the commission, has repeatedly emphasized in his interaction with this committee that the Committee on Analytics is the ultimate focus of this endeavor and all other committee's recommendations should be directed to that.

Prof Hand's Pyramid is reproduced below.



Prof Hand righty noted that researchers and policy evaluators would like to have as much granular data as possible while decision makers would like to have broad trend analysis of key performance indicators along with deeper insight on dynamic forces nehind these changes, wherever necessary. But given the vast swathe that the OSS covers, any static requirement analysis that is usually undertaken while building an enterprise Decision Support System (DSS) would be of limited use in case of official statistics. For a business enterprise requirements are generally compact and key performance indicators are not too many. For example, the core queries that a data warehouse of a supermarket chain would be required to address would be standard and a data model can be provided as an off-the-shelf product. That is why we have standard DW model as a product for commercial banks or an insurance company.

In regard to official statistics, standardization has taken place with regard to statistical workflow management system. The Generic Statistical

Business Process Model (GSBPM) has been developed by the statistical agencies of OECD countries as a by-product of their efforts to create a common metadata framework. The System of National Accounts (SNA) provides the framework for measurement of well-being of people of a nation and all nations adhere to it to a large extent to enable comparison of the state of affair of different nations. But SNA cannot be the overarching framework for the entire gamut of official statistics as many components of it are beyond the realm of financial accounting. For example, the caste composition of educational attainment of girl children is a legitimate data element that policy makers would like to monitor and track. SNA cannot be of any help in this regard. But generation of this statistics and metadata required for the same can be or rather should be carried out in accordance with a standard that is pre-determined and understood by all involved in the production of this specific statistical output. GSBPM is a work-in-progress towards that direction. Indian OSS is yet to formally adopt the GSBPM as its Statistical Workflow Management framework. However, CSO or NSSO largely manage their statistical workflow on the similar lines. Nevertheless, a formal articulation of the workflow followed by Indian statistical agencies and identification of gaps in comparison to the GSBPM standard would go a long way to streamline the existing workflow and bring them in alignment with best international practices.

GSBPM consider Metadata management as an overarching process. The Common Metadata Framework has been developed by National Statistical Institutes of various OECD countries. This framework is built upon Statistics New Zealand's approach to metadata management titled "End-to-End Metadata life-Cycle". Under this approach metadata is managed from end –to-end in the data life cycle. The importance that these national statistical authorities give to metadata management can be understood from the fact UN Economic Commission for Europe has created a metadata specific group of experts called METIS to deliberate on statistical metadata management issues.

As an example of metadata scheme followed by different countries, the metadata classification scheme followed by the Statistics New Zealand may be seen:

Conceptual Metadata : It includes concepts that underlies a specific statistics or a variable on which data is being collected(say Income), classifications, measurement units, statistical object types. The concepts can be described in a generic way or in the specific context of a particular survey or any other data collection process.

Operational metadata :This metadata type capture details about data that is relevant only from data management perspective. It would include metadata about data acquisition process – data source, data acquisition, data transformation etc.

Quality metadata: These are the metadata connected to a particular instance of a statistical object. For example, response rate of a particular survey, survey questionnaire versions etc.

Physical metadata: This metadata type includes physical characteristics of a dataset- location, server details, database details etc.

It is not suggested by this committee that Indian OSS should adopt any particular metadata architecture of any developed country. But the best practices must be studied and adopted with customization to fit Indian environment. In fact, one aspect in metadata management that has not been dealt with explicitly in the Common Metadata Framework is the notion of Semantic Interoperability or Ontological Interoperability.

Semantic Interoperability has been defined as “integrating resources that were developed using different vocabularies and different perspectives on the data. To achieve semantic interoperability, systems must be able to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands". It is obvious that building an usable and efficient Information Repository for official statistics hinges upon implementation of

required semantic interoperability among the datasets. In this context the distinction between “vocabularies” and “perspective” is fundamental and must be addressed while creating metadata management framework.

Technically “vocabularies” refers to any data object in its “as it is” form (ontology) and what can we know from the same object (epistemology). For example, let us analyze the statement by RBI’s Monetary Policy Committee- “*The projected moderation in inflation in the second half is on account of strong favourable base effects, including unwinding of the 7th CPC’s HRA impact, and a softer food inflation forecast, given the assumption of normal monsoon and effective supply management by the Government*”. If we examine this statement we encounter certain terms which are part of vocabularies: - inflation, base effects, normal monsoon, food inflation forecast, and supply management by the Government. But terms like “projected moderation”, “softer forecast” “normal” and “effective supply management” are meaningful in the specific context that these terms have been used. So to deconstruct this statement we must be told what these terms signify in this specific context. In the context of official statistics management of this distinction is of paramount importance. While ontologies are easily discoverable without any reference to user requirements, perspective is specific to a given set of related business questions that one or more datasets should be able to answer.

In regard to official statistics, data integration has multiple connotations. The High Level Group for the Modernization of Official Statistics (HLG MOS) in their publication “ A Guide to Data Integration for Official Statistics” has identified following five types of data integration relevant for official statistics.

1. Integrating administrative data with survey data and other traditional data
2. Traditional data with new data source (big data)
3. Geospatial data with statistical information
4. Macro level data with data at the macro level
5. Validating data from official sources with data from other sources.

Although all the above types of integration are relevant for creating NIDS because such integration is highly relevant for a platform that is designed to be supermarket of all official statistics that is queryable with the help of a rich metadata superstructure, it does not address the problem that is the subject matter of this section. In other words, there could multiple integrated datasets on a variety of subject matters ranging from national accounts to school enrolment, but the issue remains how all of them can be accessible through a single gateway to official statistics. The envisioned NIDS is expected provide that kind of facility to policy makers and policy evaluators. The proposed NIDS will create an integrated database system linking separate datasets with a common standard so that data can be easily correlated by leveraging the common metadata framework and a standard protocol for data exchange and data access.

The committee believes that the proposedNIDS has to be constructed through a two pronged approaches. The first approach is a federated database approach that allows end users to access data available and remaining with data providers. Such data are not physically brought to any central location and no curation of data happens by any central authority. For most of the state level data, this could be the only feasible solution. Even for data available with various central departments or regulators like SEBI, RBI, IRDA etc. this could be the preferred and most efficient approach. The second approach is to build integrated datasets as mentioned above and create, wherever possible, an analytical database to support policy makers and researchers.

For the first approach, creation of metadata must be the responsibility of the NIDS authority. Imposition of an agreed metadata framework on all data producers who are ready to provide data through the NIDS gateway is a must.

For the second approach, the requirements of specific Macroeconomic Decision Support System should be identified and created on NIDS platform. (For illustration purpose we have included two instances of the technical architecture of such data repository platforms)

Given the broad technical overview of the proposed NIDS, the Committee is of the view that it is in its merit to suggest a broad data governance architecture of the proposed NIDS in accordance with the fundamental principles of Official Statistics as adopted bythe Government of India. Thus official statistical agencies must honor “citizen’s entitlement to public information” apart from “serving the government”. Furthermore, official statistics must decide “on the methods and procedures for the collection, processing, storage and presentation of statistical data” to ensure that the official statistics is trustworthy, credible and timely.

For official statistical system, data governance architecture should incorporate, at the minimum, the following components:

1) A data management framework specifying roles, rights and responsibilities of various agencies involved in creating, capturing, processing and dissemination of data
2) A common metadata framework
3) A standard for technological infrastructure that all agencies must adhere to. Such a standard should help to optimal use of resources including hardware and software.
4) A data dissemination framework leveraging the current open data portal of the GOI, thus ensuring optimum use of existing infrastructure.
5) An organizational structure with the authority to design, review, monitor and update the above components in accordance with the acts framed in this regard by the competent legislative bodies.

Keeping in mind the federal structure of Indian political as well as statistical system the Committee suggests the following organizational structure component of the data governance architecture.

An apex coordination committee should be set up at the Centre with representatives from the states as well as MOSPI and other Ministries of GOI. This committee would be empowered to work out the details of all the above components of the data governance architecture as listed above. At the state level, a corresponding coordination committee should be set up to ensure that the standards and frameworks designed at the apex level get implemented in the respective states.

As MOSPI is the biggest consumer of data produced by all other statistical agencies, both at central and state level, it is of utmost importance that there is in place a proper mechanism for flow of data from these agencies to MOSPI with minimum manual intervention. The committee suggests that the proposed apex coordination committee should work out in consultation with NIC the details of this mechanism.

6. Recommendation for a Proof-of-Concept Project (PoC)

Since building NIDS as envisaged by the committee can be considered as a long term goal only, it is recommended that a well-spaced incremental approach should be the preferred path. The Committee took a conscious decision not to recommend any specific technical architecture for the proposed NIDS. It follows that the committee cannot but be agnostic about specific data analysis tools like R/ SAS , databases like Hadoop, Oracle Exadata, Teradata etc. Such decisions cannot precede identification of various tangible milestones that would follow when in principle decision is taken by the proposed apex coordination committee for NIDS.

Pending that decision, the Committee is of the firm opinion that "testing the water before plunging into it" should be the guiding principle while embarking on such a massive reengineering of the Indian OSS. The Committee, therefore, recommends a "Proof-of-concept" approach to building of NIDS. In this respect the Committee makes the following recommendations:

1. The NSC should identify four to five official data sources with certain desirable features- (a) data stored in an industry standard database or in a machine readable format like XBRL; b) availability of detailed metadata , at least in soft copy; (c) size of data should be significant , at least in multiples of GBs , if not in TBs. (d) data should be vital importance to policy makers and finally (e) at least one state level data.
2. A prototype NIDS based on the selected datasets should be created.
3. The proposed apex coordination committee should be entrusted with task of driving this PoC project.

4. In parallel NSC should direct the apex committeeto start building a metadata management framework in coordination CSO.
5. In the first meeting of this committee it was recommended that two pilot projects could be taken up that could help the committee frame its recommendations based on the lessons learned from the execution of these pilots. The relevant portion of the minutes of the meeting is reproduced in the Annexure III.

7. Summary and Conclusions

Official statistics is the most critical input to what Tim Holt in his 2007 prudential address to the Royal Statistical Society termed as “evidence based policy making”. Apart from catering to the information need of policy makers at governmental level, official statistics is also required for production and investment decisions of domestic and foreign entrepreneurs. Citizen at large also use official data to assess and monitor the working of their governments. To enable such decision making, official statistical system must provide required data in a manner and format that users need.

As depth and breadth of official statistics is too large to be encompassed in a single central repository with single database schema, this committee is not in favour of creating such a single central repository. The committee also believes that given the federal structure of Indian OSS, such a single central repository is not administratively feasible.

Given the above environmental constraints, the observed diversity in datasets in regard to its various attributes like storage system, data granularity, data frequency, metadata etc. can be brought under an integrating framework only by imposing standards in the data governance architecture that each statistical agency is required to adopt and implement.

These standards relate primarily to establishing a common metadata framework,prescribing a standard for technological infrastructure, a

standard protocol for data flow between statistical agencies and a common data dissemination framework. In the absence of requisite as-is position data in the above areas, the committee is proposing a proof-of concept project to gather requisite data to set the standards. NSC should establish a standing empowered committee to set the standards for creating the proposed NIDS.

8. List of Recommendations:

1) NSC should make creation of an NationalIntegrated Data System (NIDS) as a long term goal of the Indian Official Statistical System
2) NIDS is being considered as an integrating framework that would enable users of official statistics to have a single view of data available in the official statistical system irrespective of the fact that underlying databases are distributed and managed by different central and state government ministries /departments.
3) An apex coordination committee should be set up at the Centre with representatives from the states as well as MOSPI and other Ministries of GOI. This committee should have permanent status and would be empowered to set out standards for development of NIDS.
4) At the state level, a similar coordination committee should be set up to ensure that the standards and frameworks designed at the apex level get implemented in the respective states.
5) The committee is proposing an incremental approach towards building NIDS. To begin with the committee is proposing that a proof-of-concept project be taken up.
6) The apex committee should take up developing the standards that will define the integration features of the proposed NIDS in parallel to directing and monitoring of the proof-of-concept project. These are – common metadata framework, standard protocol for data flow between various state and central producers and consumers of data and a common data dissemination framework.

References

1. Economic and Social Commission for Asia and the Pacific (ESCAP) 2013 Modernization of statistical information systems Global

initiatives https://www.unescap.org/sites/default/files/Modernization-global-initiatives.pdf

2. Eurostat 2008 Conference paper on Modern Statistics for Modern Society http://ec.europa.eu/eurostat/documents/3888793/5843097/KS-RA-08-004-EN.PDF
3. Hand David J. (2017), The Value of Official Statisticsas a Public Good https://www.ine.pt/scripts/esd/presentations/David_Hand_Presentation.pdf
4. High Level Group for the Modernisation of Official Statistics, 2017-A Guide to Data Integration for Official Statistics. https://statswiki.unece.org/download/attachments/.../DIGuideDRAFTV1.0-1.pdf
5. HoltD. Tim, 2008 Official statistics, public policy and public trust
   a. in Journal of Royal Statistical Society vol171, Part 2, pp. 323–346
6. Statistics New Zealand 2015, Data Integration Manual 2nd edition. http://archive.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx
7. Telford Jenny,RooziAraghi& Peter Samson, 2015 Modernization Processes in National Statistical Offices - Transforming the Australian Bureau of Statisticshttps://www.iaos2016.ae/uploadfiles/c3df9326-7bea-4289-8d1d-c673c17b3748___jenny%20telford%20paper.pdf

Constitution of the Committee on Analytics

1.	Prof N.L. Sarda, Deptt. Of Computer Science, Indian Institute of Technology, Mumbai Email : nls@cse.iitb.ac.in	Chairman
2.	Dr. K.R. Murali Mohan, Deptt. Of Science & Technology,	Co-Chair

	Ministry of Science & Technology, Email :krmm@nic.in	
3.	Dr. A.C. Kulshreshtha, Former Addl. DG, CSO (NAD) 208 E, MIG Flats, Rajouri Garden, New Delhi 110 027 Tel.: 011-25972191 Email: ackulshreshtha@yahoo.com	Non Official Member
4.	Prof. Pulak Ghosh, Presidency University, Kolkata Mobile 9742065806	Non Official Member
5.	Shri Pramod Varma, Former Chief Architect, UIDAI Technology Centre, NTI Layout, Tata Nagar, Kodigihalli, Bengaluru 560 092 Email: pramodkvarma@gmail.com	Non Official Member
6.	Dr. Ashok K Nag, Director, Centre of Excellence in Analytics, NM University, Mumbai Email: ashok.nag@gmail.com	Non Official Member
8.	Shri Pravin Srivastava, DDG (Directorate General of Employment & Training), M/o Labour & Employment	Member
9.	ADG, NAD, Central Statistical Office	Member
10.	Representative from National Informatics Centre	Member
11.	Advisor in Charge of Information Management, DSIM, RBI	Member
12.	ADG, DPD, National Sample Survey Office	Member
13.	Director General, Directorate General of Commercial intelligence & Statistics	Member
14.	Representative from Ministry of Finance	Member
15.	Representative from Ministry of Agriculture	Member
16.	Representative from Ministry of Commerce & Industry	Member
17.	Director, DES, Maharashtra	Member
18.	Director, DES, Karnataka	Member
19.	Director, DES, Uttar Pradesh	Member

20.	Director, DES, Odisha	Member
21.	ADG, Computer Centre	Member Secretary

ANNEXURES

Annexure 1: Metadata; A suggestion by one National Statistical Institute

Annexure 2: Examples of Official Statistical Data Repository Architectures from Different Countries:

Annexure 3: Excerpt from the Minutes of the First Meeting of Committee on Analytics Held on 16.01.2017

Annexure 4: **Minutes of the 2nd Meeting of the Committee on Analytics held on 10th May, 2018 at IIT, Mumbai**

Annexure 1

Metadata – A suggestion by one National Statistical Institute

The following metadata are vital for end users of statistical metadata and data:

• Availability of statistical outputs;

• Metadata related to the statistical outputs (metadata and data concepts and definitions, classifications, aggregations, statistical and evaluation methods, terminology, history, etc);

• Metadata about quality (e.g. explanatory notes);

• Access to microdata;

• Time series;

• Updating procedures;

• Statistical revisions;

• Responsibility for individual statistical outputs;

• Links to other information systems both national and international;

• Confidentiality;

• Planned changes in statistical outputs;

• Content related standards, both national and international;

• Outcomes from statistical analysis on users feedback;

• Rules for searching, accessing and downloading statistical metadata and data from output databases;

• Technological standards relevant for extraction and transfer of data and metadata;

• Information about software and other tools supporting search, retrieval and downloading of metadata and data;

• Users training possibilities;

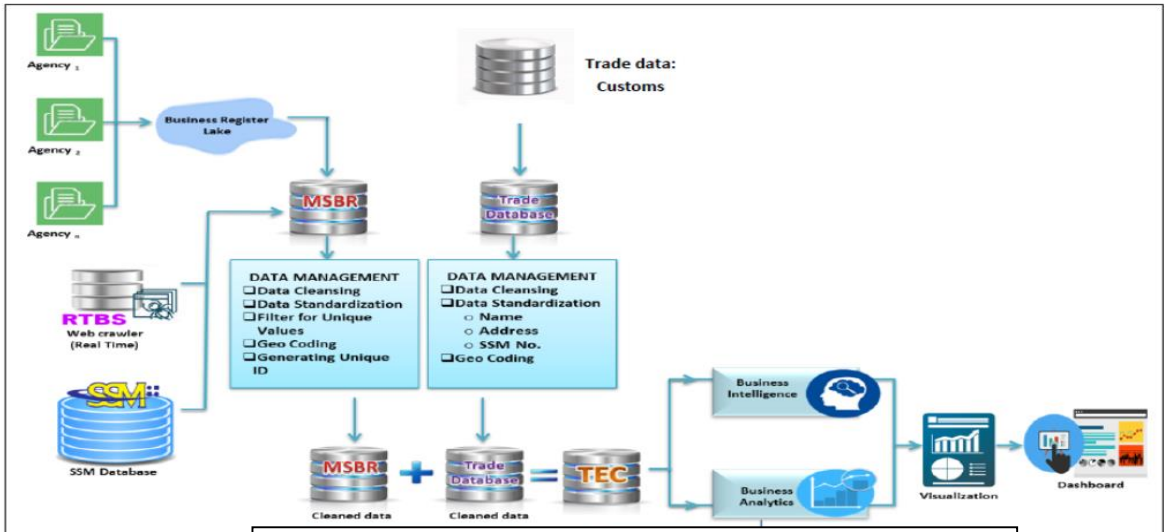
• Metadata based services such as classification coders and metadata mappings that other producers and users of statistics can apply.

See: United Nations Economic Commission for Europe: A Common Metadata Framework: Part A:Statistical Metadata in a Corporate Context 2009Retrieved on 29th May 2018

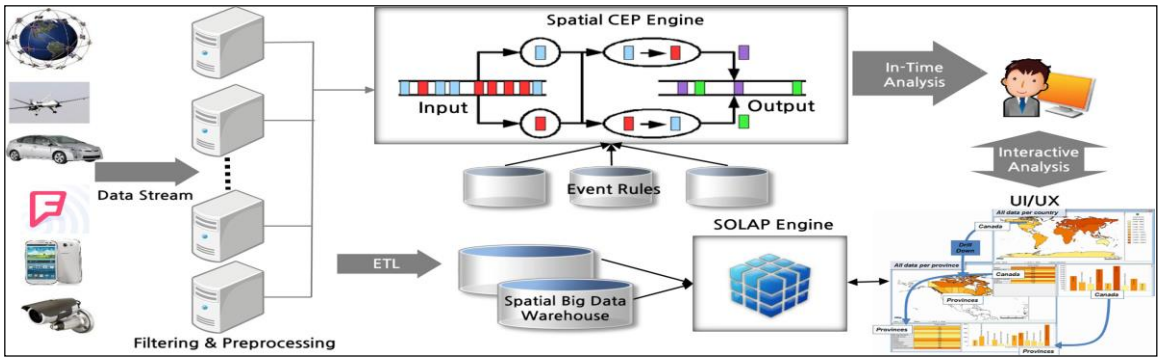
http://www.unece.org/fileadmin/DAM/stats/publications/CMF_PartA.pdf

Annexure: 2

Examples of Official Statistical Data Repository Architectures from Different Countries:



Examples of Spatial Big Data warehouse Platforms



CEP = complex event processing, SOLAP = spatial online analytical processing.
ETL = extract, transform and load, UI/UX = user interface/user experience design.

Annexure 3

Excerpt from the Minutes of the Meeting of Committee on Analytics Held on 16.01.2017 (Monday) at Ahmedabad, Gujarat under the Chairmanship of Prof. N.L.Sarda.

1. A few departments to do a case study on sharing their data and metadata and usage scenarios to host their data on CSO's future platforms consisting of the main components as Metadata repository, Data System (may be combinations of files, databases or warehouses), portal systems and stat tools such as R. These departments could be from among National accounts, trade, DGCI, NSSO, Education and Agriculture.

2. To use the models and platforms of Gujarat Stat Dept to combine and merge data from Gujarat with two more states using their 2800+ data items and subject areas as the metadata backbone, and developing similar portals and stat reports as theirs but covering data aggregated from 3 states. Also, using their village and Taluka Profiling System (with important extensions suggested yesterday to proactively communicate stat findings with important stakeholders. In order to do this pilot effectively, Gujarat may be requested to lead this pilot (even possibly using the same technical teams of their solution vendor) so that not only the pilot can be completed in time, but the validation for integration driven by metadata is demonstrated. This pilot can use R in place of the current commercial tool, and can also attempt to put all metadata in a tool that will support querying of metadata as well as the canned reports. CSO may provide funding support (to Gujarat) for carrying out this pilot.

ANNEXURE 4

Minutes of the 2nd Meeting of the Committee on Analytics held on 10th May, 2018 at IIT, Mumbai

1. The meeting of the Committee on Analytics was held on 10th May, 2018 under the Chairmanship of Prof. N L Sarda, IIT, Bombay at IIT Mumbai Campus. Dr. R B Barman, Chairman of the National Statistical Commission was also present in the meeting. The list of participants of the meeting is attached as Annex -1.

2. At the outset of the meeting, Shri B N Tiwari, Member Secretary of the Committee welcomed the Chairman of the Committee, and other distinguished members of the committee. He accorded special welcome to Dr. R B Barman, Chairman of the National Statistical Commission who took pain and spared times from his busy schedule to make himself convenient for the meeting. Shri Tiwari briefed the committee about the background of the committee and its objectives. He highlighted that considering the vitality of the committee for the Indian statistical system. Originally the distinguished personalities who have exemplary achievements and remarkable contributions in the related fields were opted as non official members of the committee. They are Dr. K R. Murali Mohan, DST, Dr. A.C. Kulshreshtha, Ex-ADG, MoSPI, Prof. Pulak Ghosh, Presidency University, Kolkata, Shri Pramod Varma, Former Chief Architect, UIDAI technology centre, Bangaluru and Dr. Ashok Nag, former, Adviser, RBI. Later, Prof. S. Rajagopalan, IIIT, Bengaluru was inducted in the committee to benefit from his expertise in GIS and BI. However, their expertise/views could not be effectively utilised. He further summarized discussions held in the first meeting of the committee in Gujarat as under :

    (i) Members and CSO to provide brief summary of statistical system of their organisations indicating domain areas, systems, tools, internal data warehouses, primary sources of data etc.
    (ii) CSO to prepare a report on best practices outside India
    (iii) Prepare a feedback from data users on CSO data/ reports on adequacy of statistical data and reports

(iv) To initiate a few pilot studies that help in preparing a roadmap as well as a blue print for strengthening systems as well as data in CSO.

He also informed that DSDD wrote several times to all members who participated in the first meeting but couldn't receive any action taken reports or inputs from them except from RBI and Dr. Ashok Nag. This was not appreciated by the Chairman, NSC. He solicited the active participation and contributions of each member so that the Committee can deliver as per expectations.

3. Furthering the agenda of the meeting Dr. Barman shared his experiences while working on the Date Warehouse Project of RBI. **Dr Barman cautioned that without proper preparation, we should not jump into the project. He stated that while associated with similar projects for over 15 years, he is of the view that while we need to get the project done by the vendors, we need to be on driver's seat for clearly drawing up system requirements.** We should also have appreciation level knowledge on systems development for Analytics. These abilities are essential to avoid bickering and also setting clear targets on deliverables at each stage of the project. We may start the project with a pilot study for hands on experience which may be based on at least one easily manageable area. The area may be selected considering ease on data availability and versatile on dimensional features for classification of data. On the basis of experiences and feedbacks from the pilot studies, the internal team may be well prepared to drive the project. He expressed his deep concern that the Indian Statistical System is still at very nascent stage in adoption of ICT for digital transformation. **There are issues on data quality, consistency, coherence, timeliness and transparency which cannot take a back seat any longer.** Our policy need to be focus on shared vision of national development priorities, sectors and strategies with the active involvement of the states in the light of national objectives. We need to support formulation credible plans and their monitoring at all levels of governance as set in the vision document of NITI Aayog. This will give

the ability to evaluate governance much more effectively. The system should support knowledge economy, as public good.

4. Thereafter, he made a presentation on Analytics - Building Blocks before the committee. He stressed that technology can help us in having a robust system through bottom up approach on collection of data and top down approach for development of systems using Analytics. In addition to conventional output, data visualization, automated MIS for all levels of governance and 'what if' analysis can give us much deeper insight for formulation of policy, monitoring and evaluation of progress. He quoted Irma Adelman who stated that "development should be analysed as a highly multifaceted, nonlinear, path-dependent, dynamic process involving systematically shifting interaction patterns that require changes in policies and institutions over time." He referred a case study of UK under title "The Office for National Statistics Statistical Modernization Programme: What went right? What went wrong?" by Stephen Penneck. Penneck mentioned that pressures to operate more efficiently, respond more rapidly to changing user demands, exploit data more effectively and improve statistical quality have led a number of statistics offices to seek to modernize their statistical systems through adopting an information technological environment, using standard tools and processes across statistical systems, with common business processes driven by metadata. The UK's office for National Statistics embarked on such a programme of statistical modernization in 2001. Together the Design Authority and IT strategy provide clear direction for modernisation. While it has not achieved all its goal, there have been many achievements. Penneck paper reviewed the programme and set out what was not achieved. His paper would help us to draw lessons which will be relevant for the future modernization of statistical offices.

5. Mr. Barman also highlighted Beans Committee Report-2016. Mr. Charlie Bean, a former deputy governor of the Bank of England, outlined as under transform UK economic statistics and to fully capture all the activity in the economy.

(i) Assess the UK's future (economic) statistics needs, in particular relating to challenges of measuring the modern economy ('Needs');
(ii) Assessing the effectiveness of the ONS in delivering those statistics, including the extent to which ONS makes use of relevant data and emerging data science techniques ('Capability');
(iii) While fully protecting the independence of U.K. National Statistics, consider whether the current governance framework best supports the production of world-class economic statistics ('Governance').
(iv) The Review was prompted by the growing difficulty of measuring output and productivity accurately in a modern, dynamic and increasingly diverse and digital economy.
(v) Greater analytical capability, both in economic understanding and the ability to handle and interrogate large data sets.

6. On a technology front, he was impressed by the concept of "Spatial Big Data Warehouse Platforms" for which a flow chart was presented as illustration. He suggested a focused discussion on major building blocks as under:

Data Capture - Whether it should be latest concept of Data Lake or any other along with data ingestion. Considering structured nature of data used by our statistical system we need to find an optimal solution. **Considering highly decentralised nature of our system, whether a cloud based clustered approach with virtualisation will be appropriate for smooth integration following certain standard needs to be considered. It may be desirable to keep in view the possibility of horizontal scaling and tools for parallel processing**.

Conformed dimensions validated by Master Data Management (MDM), Multi Dimensional Data Base (MDDB) are helpful for striving for single version of the truth. Open source tools like R and Python along with other user preferred tools may be required for deep analysis, data visualisation tools, dashboards (MIS).

XML defined data/JSON (Java Script Object Notation) - XBRL and SDMX for data exchange

Sandbox - for exploratory data analysis and discovery process.

Data Governance and Access Control

7. In order to Plan the Project, effectively, he suggested as under:

(i) Clear understanding of conceptual foundation of decision making for producing statistics, business requirement. clear understanding of business goal helps in designing Schema in quickest time.

(ii) For BI analysis, a step-by-step methodology is needed to organise the activities and tasks involved with acquiring, processing, analysing and maintaining data.

(iii) Major steps - Clearly defining data requirement, Data Identification, Data Definition and Metadata, Data Acquisition and Filtering, Data Extraction, Data Validation and Cleansing, Data Aggregation and Representation, Data Analysis, Data Visualisation.

He concluded his presentation once again re-iterating that the prototype/ pilot project on Data Warehouse should be invaluable in going for the Project. In his view the present state of technology provides enough flexibility to use RDBMS and HDFS for **data and parallel processing for quick processing when the volume is likely to grow very high**. As official statistics follow well defined methodology a Data Warehouse using Big Data Technology appears relevant for large scale processing. He also mentioned that a central Design Authority, the way UK has gone, appears very relevant for such a huge project to benefit from, progress according to plan and succeed.

8. The presentation was appreciated by all members. At the technology front, Chairman of the committee, Prof. N L Sarda, expressed his reservations. He was of the view that we should not put the cart before the horse. We need not suggest or fix technology at the outset as the technology is continuously growing and changing. We should understand the requirements in general and domain-independent manner and map them on a conceptual architecture that can stand test of time and be flexible enough to allow use of emerging technologies and solutions/tools. The

problem may be addressed by suitable technology of the time and multiple technologies can co-exist as long as they permit inter-operability. This view was also appreciated by other members of the committee. He also resisted that the job of this committee does not include supervising the pilot project. He stated that it is not necessary to understand nature or domains where the Department is handling today's data. A general understanding as given by the Chairman NSC in his multiple presentations is adequate. He suggested that the committee should give its recommendations on the approach and flexible data analytics and integration platform that will be not tied to a specific technology.

9. Thereafter, the comments from other members were invited. These are summarized as under:

(i) Dr Ashok Nag, Member of the committee mentioned that this committee on Analytics needs to look into the entire life cycle of production of statistical outputs and suggest the kind of analytics that can be used at each stage of this life cycle. For example, at the data sourcing stage, we can explore how “Big Data” can be used. Similarly, efforts are underway by some official statisticians to make use of Artificial Intelligence techniques like Neural net in automatic editing of collected data. At the data user and data dissemination end many statistical authorities are providing facilities to create user defined reports based on data available for public dissemination. If we want to suggest the proto type project that may be done. But, this should not be only recommendation of the committee. We also need to deliberate on the specific recommendations that this committee should make. Our report should explain the rationale of its recommendations and its contextual background. We need not recommend technology as technology is a tool. It is not possible for this committee to recommend a generic tool unless the committee knows the specific context in which the tool is to be used.. It would be presumptuous on our part to claim that we are in know of all the requirements of all users of official statistics. It

would be easier for the committee to provide a Framework for technology mediated informational platform that would facilitate use of analytics.

(ii) Prof Sarda, Chairman of the Committee, told that we need to develop framework for sharing of data and integration with various source agencies. He gave the example of NSDI and **data.gov.in**. He also informed that he is also a part of the committee which is entrusted to develop National Data Registry which will facilitate data integration and usage across various data generating and consuming agencies. Today, data of various ministries are not integrated. The Union Government is developing a National Data Registry (NDR) to compile and serve metadata of different agencies and re-engineer the feature data set for improving their reuse. The registry will also serve as a source of authenticated information. The Department of Science and Technology (DST) will be the nodal coordinating agency of the NDR. NDR will require all agencies state, private and academic, collecting and storing geospatial data to provide details of data they store and allow access through services. The registry aims to create a catalogue that will prevent duplication of data sets and to help users locate the right agencies to source information. It will be a metadata repository to only inform about nature of the data a service provider has. Here this committee is not recommending any technology, but only an enterprise architecture. Similarly here also this committee will not recommend any technology. We can recommend a similar generic architecture. He suggested that data.gov.in is actually very agile platform which enabled source agencies to publish their data on the platform easily. In case of a Data Warehouse, it is very rigid and also involves data cleaning, sorting and integrating various types of information upfront, restricting easy usage of data. Integration of entire statistical information at one platform may not be possible. Rather, he emphasized that we need to develop federation where each source agency has its

own data centre and that can share data services to the central platform. Actually we need a very flexible platform.

(iii) Shri Pravin Srivastava, Addl. DG, NAD informed the committee that decision to develop Data warehouse was taken in way back 1999. We lost so much time. NSC has constituted two very important committees viz. Committee on Analytics and Committee on Online Reporting System. There are commonalities in both committees one in part of technology and other in the way integration is required. But the Ministry has to implement the National Data warehouse for this we have already prepared RFP and pending with IFD for approval. We have got the structure approved. He suggested that this committee may recommend the ways on which this project can be implemented immediately. If the report is like text book that may not have any relevance as there are already so much textbooks in the market. NSC chairman responded that statistical system is very huge and we can't have a single broad bust to solve the problems. We need to solve the problem with a small start with deliverables. RFP is too big to deliver. **He cautioned that we need to be very clear in our approach. He stated that any good vendor would go mad if asked to integrate entire statistical system. Integrating entire statistical information on one platform is not possible, in fact next to impossible.** Prof Sarda also mentioned that any review or recommendations on the present Data Warehouse RFP is outside the scope of this committee.

(iv) Shri T K Saha, Addl. DG DPD, mentioned that the data are being produced by various agencies at various levels of Governance. MOSPI is producing data on EC, NSS, ASI, etc and State Govts. are also producing the data on various subjects. Integrating and mapping of entire information is very difficult. This committee may deliberate on this.

(v) Shri Sanjiv Basu, Director, DGCIS, informed that DGCIS is developing a Data Warehouse of the statistics produced by it and it is pending in the ministry for approval.

(vi) Shri Anujit, Adviser RBI informed that in comparison to RBI data warehouse, the proposed data ware house is too big. The technology has also advanced since then. But, we need to have a very flexible platform to accommodate various different types of data.

(vii) Smt. Mamta Saxena, Statistical Adviser, M/o Agriculture supported the ideas of Prof. Sarda and she said that **we need to have a very flexible platform which should be accommodating to the federal structures of the Governance. She advised that each source agency may have their own data centre and these data centres may be mapped as per central requirements.**

(viii) Shri B.N. Tiwari, ADG, DSDD stated that considering the huge dimensions of the Indian statistical system, as discussed, it is very difficult to integrate the entire system at one common platform. **He supported the idea that each source agency may have their own Data Centre and online reporting system.** This Committee may recommend some standard protocols for online data collection and also data sharing, dissemination and integrating to be observed by each source agency i.e., Ministries/Departments of the Central/State Governments/PSUs which may later on be integrated with the centralised system, on need basis.

10. Thereafter, the Chairman of the committee summarized the discussions and stated that this would be the last meeting of the Committee. In order to prepare a draft report, Chairman informed that he will prepare some bullet points outlining structure of the report and our recommended generic architecture, which he would send to Member Secretary. Member Secretary may elaborate these points and prepare a draft report. The draft report may be sent to Dr Ashok Nag for further elaborations. There after this draft report will be circulated to all members for taking their final comments by 20th May, 2018. The final reports of the committee may be ready by first week of June, 2018.

The meeting concluded with thanks to the Chair and members of the Committee. Chairman specially thanked Dr. R.B.Barman who spared time from his busy schedule to participate in the meeting. The Committee has immensely benefited by his presence and presentation.

****