REPORT

OF

THE COMMITTEE ON DATA MANAGEMENT

GOVERNMENT OF INDIA

NATIONAL STATISTICAL COMMISSION

MINISTRY OF STATISTICS AND PROGRAMME IMPLEMENTATION

NEW DELHI

CONTENTS

Page Number

Constitution, Terms of Reference and Members' List (i)

Acknowledgements (vii)

List of Abbreviations (ix)

Introduction, Executive Summary, and Main observations (xviii)

Chapter - 1 Indian Statistical system, Challenges in Official Statistics & data management and remedial measures

1. The importance of official statistics 1
2. Current scenario of National Statistical System in India 3
3. Need to know and right to share 4
4. Issues of decentralized system 5
5. Vertical and horizontal issues 7
6. Priority 9
7. Inter operability 9
8. Need for assessability and communicability 10
9. Need for scanning of published data before conversion into defined data format 11
10. Non-official data providers 12
11. Consistency in definition and standards 13
12. Draw upon insights into problems 13
13. Technological and technical issues 15
14. Training intervention/ capacity building 17
15. Technology versus Applications 19
16. Legal framework for providing data 19
17. Setting up of Data Management Division 20

Chapter - 2 Conceptual framework on Data management

1. Data Management 22
2. Data governance 25
3. Data Architecture, Analysis and Design 26
4. Database Management 29
5. Data Security Management 30
6. Data Quality Management 32
7. Reference and Master Data Management 34
8. Data Warehousing and Business Intelligence Management 36
9. Document, Record and Content Management 39
10. Meta Data Management 40

11. Contact Data Management 42
12. Data Movement 44
13. Statistical Data Metadata Exchange (SDMX) 44
14. Secured Data Centers 46

Chapter -3 Recommendations:

1. Short term (Implemented within six months) 50
2. Long term (Implemented within one to two years) 52

Annexure Minutes of the meetings of the Committee on Data Management

1. 1.a Minutes of the first meeting of the Committee 57
   1.b Minutes of the second meeting of the Committee 63
   1.c Minutes of the Third meeting of the Committee 72
2. Current national statistical system on key sectors 78
3. Data Dissemination Policy of Government of India 112

Constitution, Terms of Reference and Members' List

No.8(64)/2010-NSC
Government of India
Ministry of Statistics and Programme Implementation
National Statistical Commission Secretariat
II Floor, Sardar Patel Bhawan,
Sansad Marg, New Delhi-110001
The 30th July 2010

ORDER

Sub: Constitution of professional committees by the NSC

The issue of constitution professional committee to assist it on various technical issues were under the consideration of the National Statistical Commission (NSC) for quite some time. It has now been decided by the NSC to constitute eight professional committees. The details of composition and terms of reference of each committee are given in the Annexure. The Committee on Statistics of Agriculture & Allied sector will have tenure of twelve months. All the remaining Committees will have tenure of six months. The Committees shall accordingly submit their reports to the NSC.

2. With the approval of the Chairman of the NSC, each of the Committees may also enlist the assistance of subject matter experts within and/or outside the Government and may co-opt them as members according to necessity.

3. The expenditure on TA/DA of the official members will be borne by their respective Ministries/Departments/Organizations. Each of the non-official Members would be entitled for a sitting fee of Rs.1, 000/- per day for attending meetings. They will be eligible to travel by air in executive class or by rail in air-conditioned first class while undertaking tours in connection with the meetings of the respective Committees. They will also be entitled to TA/DA on tours as admissible to a Joint Secretary to the Government. Besides, they will also be entitled to transport or transport charges for local travel for attending the meetings of the respective Committees.

4. Secretariat support to the Committees would be provided by the Central Statistical Organisation. The expenditure on conducting the meetings of the Committees and on payments/reimbursements made to the non-official Members will, under the relevant heads, be de bitable to the budget allocated to the NSC under the non-plan grant of the Ministry of Statistics & Program Implementation (MOSPI).

5. This issues with the concurrence of IFD vide Dy. No. 512/B&F dated 29th July, 2010.

6. This order comes into effect from 1st August 2010.

Sd/-

(M.V.S. Ranganadham)
Dy. Director General
Telefax: 011-23367128
Mob: 919818878155
E-mail: nsc-secretariat@nic.in

Copy to:

1. Chairman & Members of the eight Committees
2. Chief Secretary, Govt. of Uttar Pradesh/Maharashtra/Andhra Pradesh/West Bengal/Assam/Himachal Pradesh/Bihar/Mizoram/Punjab/Tail Nadu
3. Secretary, M/o Finance, Deptt. of Economic Affairs, New Delhi with a request to nominate a suitable officer in the Committee No. II, III.
4. Secretary, Planning Commission, New Delhi.
5. Secretary, Dept. of Information Technology, New Delhi
6. Secretary, Dept. of Industrial Policy and Promotion, New Delhi.
7. Secretary, Dept. of Consumer Affairs, New Delhi.
8. Secretary, M/o Corporate Affairs, New Delhi.
9. Secretary, M/o Health & Family Welfare, New Delhi with a request to nominate a suitable officer in the Committee no. VI.

10.Secretary, M/o Human Resource Development, New Delhi with a request to nominate a suitable officer in the Committee no. VI.
11. Secretary, M/o Social Justice & Empowerment, New Delhi with a request to nominate a suitable officer in the Committee no. VI.
12. Secretary, M/o Women & Child Development, New Delhi with a request to nominate a suitable officer in the Committee no. VI.
13. Secretary, M/o Minority Affairs, New Delhi with a request to nominate a suitable officer in the Committee no. VI.
14. Secretary, NCERT, New Delhi with a request to nominate a suitable officer in the Committee no. VI.
15. Dy. Governor (Dr. Subir Gokam), Reserve Bank of India, Mumbai with a request to nominate a suitable officer in the Committee no. VI.
16. Director General, Central Statistical Office.
17. Director General, National Sample Survey Office.
18. Addl. Director General, NSSO (FOD Hqrs), New Delhi
19. Addl. DG, NSSO (SDRD), Kolkata
20. Addl.DG, NSSO (DPD), Kolkata.

IV. Committee on data management

Composition:

1.	Shri Suman K.Bery, Member, NSC E-Mail: sbery@ncaer.org	Chairman
2.	Shri Deepak Mohanty, Executive Director, RBI, Mumbai E-Mail: dmohanty@rbi.org.in	Member
3.	Dr. K Kanagasabapathy, EPW Foundation	Member
4.	Dr. Rajesh Shukla, NCAER, New Delhi E-Mail: rkshukla@ncaer.org	Member
5.	Director, RBI Institute for Development & Research in Banking Technology, Hyderabad	Member
6.	Representative of NIC	Member
7.	Representative of CMIE	Member
8.	Representative of ISI	Member
9.	Addl. DG, CSO (NAD)	Member
10.	Representative of NIC	Member
11.	Addl. DG, NSSO (DPD)	Member
12.	DDG, CSO (IS Wing)	Member
13.	IT Coordinator, MOSPI	Member
14.	DDG, Computer Centre	Member Secretary

Terms of reference:

Data Management using developments in information technology and dissemination co forming to International Standards Data Management using developments in information technology and dissemination conforming to International Standards

Sd/-

(M.V.S. Ranganadham)
Dy. Director General
Telefax: 011-23367128
Mob: 919818878155
E-mail: nsc-secretariat@nic.in

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

II Floor, Sardar Patel Bhawan,
Sansad Marg, New Delhi-110001
The 12th August 2010

CORRIGENDUM

Sub: Constitution of professional committees by the NSC

In partial modification of this Ministry's Order of even no. dated 30th July 2010 on the above subject, it is hereby ordered that Dr. S. Durai Raju, Dy. Director General, CSO (NAD) would be the Member Secretary of the Committee (No. IV) on Data Management in place of the DDG, Computer Centre. It is also ordered that DDG, Computer Centre would be a Member of the Committee. The details of the revised composition of the Committee No. IV are enclosed for ready reference.

2. This issue with the approval of the Secretary.

Sd/-

(M.V.S. Ranganadham)
Dy. Director General
Telefax: 011-23367128
Mob: 919818878155
E-Mail: nsc-scretariat@nic.in

Copy to:-

1. Chairman & Members of the Committee on Data Management
2. Secretary, Dept. of Information Technology, New Delhi.
3. Director General, Central Statistical Office.
4. Director General, National Sample Survey Office.
5. Addl. Director General, NSSO (FOD Hqrs), New Delhi.
6. Addl. DG, NSSO (SDRD), Kolkata
7. Addl. DG, NSSO (DPD), Kolkata
8. Addl. DG, CSO (NAD/SSD/ESD/NASA/CAP), New Delhi
9. Director, Indian Statistical Institute, Kolkata with a request to nominate a suitable officer in the Committee.
10. Dy. Director General, CSO (IS Wing), Kolkata
11. Dy. Director General, Computer Centre, R.K. Puram, New Delhi.
12. Managing Director, CMIE, Mumbai
13. Joint Secretary (Admn), MOSPI, New Delhi
14. Director & HOD, MOSPI, New Delhi

15. Director (IFD), MOSPI, New Delhi
16.Pay & Accounts Officer, MOSPI, New Delhi
17.Admn.I Section, MOSPI, New Delhi
18.General Section, MOSPI, New Delhi
19.Cash & Accounts Section, MOSPI, New Delhi
20.Hindi Section for Hindi version

Copy also for information to:-

1. Chairman & Members of the NSC
2. PPS to Secretary (S&PI)
3. PS to AS (S&PI)
4. PPS to AS & FA

Sd/-

(M.V.S. Ranganadham)
Dy. Director General

IV. Committee on data management

Composition:

1.	Shri Suman K.Bery, Member, NSC E-Mail: sbery@ncaer.org	Chairman
2.	Shri Deepak Mohanty, Executive Director, RBI, Mumbai E-Mail: dmohanty@rbi.org.in	Member
3.	Dr. Kanagasabapathy, EPW Foundation E-Mail: ksabapathy@epwrf.res.in	Member
4.	Dr. Rajesh Shukla, NCAER, New Delhi E-Mail: rkshukla@ncaer.org	Member
5.	Director, RBI Institute for Development & Research in Banking Technology, Hyderabad	Member
6.	Representative of NIC	Member
7.	Representative of CMIE	Member
8.	Representative of ISI	Member
9.	Addl. DG, CSO (NAD)	Member
10.	DDG, Computer Centre	Member
11.	Addl. DG, NSSO (DPD)	Member
12.	DDG, CSO (IS Wing)	Member
13.	IT Coordinator, MOSPI	Member
14.	Dr S Durai Raju, DDG, CSO (NAD)	Member Secretary

Acknowledgements

Acknowledgements

When the National Statistical Commission asked me to Chair the professional "Committee on Data management", to review and deliberate upon the issues relating to data management in Indian Statistical system at par with International standards and provide necessary recommendations, I was glad to accept the responsibility. I was convinced that there was need to take an overall view, highlighting the importance of linkages between various data sources, integrating the data sources using Information Technology(IT) tools and offering a generally consistent approach to bring out timely, reliable, credible and user friendly data sets through a single window system. The purpose of the report is to catalyse debate as well as action, some immediate and some over time, to bring forth the conceptual frame work on Data Management, optimum use of Information Technology in integrating and harmonizing the available data sets in various key sectors through single window system, briefing some of the challenges likely to be faced in generating, compiling, analysing, maintaining and disseminating quality data both at micro and macro (aggregate) level for the user community.

I start my acknowledgements by thanking Shri Deepak Mohanty, Executive Director, RBI and his team of officers for the proactive role, coordinating with the Members of the Group and hosting second meeting of the Committee, making available requisite important inputs for banking sector statistics and enlightening the Members of the Committee on various issues relating to the implementation of data warehousing solutions in banking sector.

Shri Ashish Kumar, Additional Director General, NAD, CSO and all other members of the Committee took on the additional burden of attending the meetings cheerfully and shared their experiences. Many members have also provided valuable inputs on data management and dissemination issues in their subject fields which are very helpful in making this report. I thank profusely all the Members for their valuable participation and contribution.

The Secretariat of National Statistical Commission has also been invaluable help to the working of this Committee, for hosting many of the meetings and providing logistical support.

My special thanks are due to Dr. S Durai Raju, DDG (NAD), Member Secretary of the Committee for his enthusiasm, involvement and active participation in the Committee and also for giving final shape to this Report.

We held three formal meetings and many informal discussions. This has been truly a collective effort, and I would like to express my sincere gratitude to all the Members, Special Invitees, RBI, CSO and the staff of the NSC.

Suman Bery

National Statistical Commission constituted a Professional Committee on Data Management vide its order dated 30th July, 2010 and revised order dated 12-th August, 2011 under the chairmanship of Shri Suman Bery, Member of National Statistical Commission. The Committee deliberated the subject in its three meetings held on 27th September, 2010 in New Delhi, on 18-th February, 2011 in Mumbai and 19th April, 2011 in New Delhi respectively.

We, the Members of the "Committee on Data Management", hereby submit the Final Report.

Shri Suman Bery,
Member, NSC
Chairman

Shri Deepak Mohanty,
ED, RBI
Member

Shri Ashish Kumar,
ADG(NAD), CSO
Member

Dr K Kanagasabapathy,
Director, EPW Foundation,
Member

Dr Rajiv Mehta,
ADG(DPD), NSSO,
Member

Shri P C Mohanan,
DDG(CC), MOSPI,
Member

Sh M V S Ranganadham,
DDG (NSC), MOSPI
Special Invitee

Shri B. Sambamurthy
Director, IDRBT,
Member

Prof. Smarajit Bose,
ISI, Kolkata,
Representative of ISI

Shri Bimal K. Giri,
CSO (IS Wing), Kalkota,
Member

Mr Rohit Sabherwal,
CMIE,
Member

Shri Sunil Jain,
Senior Tech. Director, NIC,
Member

Dr S Durai Raju,
DDG(NAD), CSO,
Member Secretary

List of Abbreviations

ADG	Additional Director General
ADO	ActiveX Data Objects
AE	Advance Estimates
AIES	All-India Educational Survey
AIITS	All India Income Tax Statistics
ANCOVA	Analysis of covariance
ANOVA	ANalysis Of VAriance
ANSI	American National Standards Institute
ASTM	American Society for Testing and Materials
AYUSH	Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy
BCP	Business continuity planning
BCRP	Business continuity & Resiliency planning
BI	Business intelligence
BIA	Business Impact Analysis
BIS	Bank for International Settlements
BSR	Basic Statistical Returns
CACP	Commission for Agricultural Costs and Prices
CAPE	Crop Acreage and Production Estimates
CAPI	Computer Aided Personal Interview
CATI	Computer Aided Telephone Interview
CBDT	Central Board of Direct Taxes
CBEC	Central Board of Excise and Customs
CBS	Consolidated Banking Statistics
CCEA	Cabinet Committee on Economic Affairs
CCE	Crop Cutting Experiments
CDA	Confirmatory Data Analysis

CEA	Central Electricity Authority
CIFRI	Central Inland Fisheries Research Institute
CIN	Corporate Index Number
CMW	Common Warehouse Metamodel
COBOL	COmmon Business-Oriented Language
COCSO	Central and State Statistical Organisations
COCSSO	Conference of Central and State Statistical Organisations
CPI	Consumer Price Index
CPI(AL)	Consumer Price Indices (Agricultural Labourers)
CPI(IW)	Consumer Price Indices (Industrial Workers)
CPI(UNME)	Consumer Price Indices (Urban Non-Manual Employees)
CPU	Central Processing Unit (in Computer)
CRR	Cash Reserve Ratio
CSO	Central Statistical Office
CWWG	Crop Weather Watch Group
DAC	Department of Agriculture & Cooperation
DARE	Department of Agricultural Research and Education
DBA	Database Administrator
DBMS	Database Management System
DCI	Dental Council of India
DCSSI	Development Commissioner of Small Scale Industries
DES	Directorate of Economics & Statistics
DESMOA	Directorate of Economics and Statistics, Ministry of Agriculture
DGCIS	Directorate General of Commercial Intelligence and Statistics
DGE&T	Director General of Employment & Training
DGET	Director General of Employment and Training
DGHS	Directorate General of Health Services
DM	database modeling

DM	Data Mart
DMBOK	Data Management Body of Knowledge
DMS	Document Management System
DPEP	District Primary Education Programme
DSD	Data Structure Definition
DSS	Decision Support System
DVD	Digital Versatile Disk
DW	data warehouse
DWBI	Data Warehousing and Business Intelligence
EARAS	Establishment of an Agency for Reporting Agricultural Statistics
EC	Economic Census
ECB	European Central Bank
ECM	Enterprise Content Management
ED	Executive Director
EDA	Exploratory Data Analysis
EDA	Electronic Design Automation
EDI	Electronic Data Interchange
EDMS	Electronic Document Management System
EDRMS	Electronic Document and Records Management System
EEZ	Exclusive Economic Zone
EII	Enterprise Information Integration
EMI	Employment Market Information
EMIP	Employment Market Information Programme
ENVIS	Environmental Information System
ERM	Electronic Record Management
ETL	Extract, Transform and Load
EUS	Employment-Unemployment Surveys
FASAL	Forecasting Agricultural Output Using Space, Agro Meteorology and Land

FOD	Field Operations Division
Fortran	FORmula TRANslation
FSI	Forest Survey of India
GCES	General Crop Estimation Survey
GDP	Gross Domestic Product
GNP	Gross National Product
GVA	Gross Value Added
H&FW	Health & Family Welfare
HII	Health Information of India
HIPAA	Health Insurance Portability and Accountability Act
HIS	Horticulture Information Systems
HMIS	Health Management information System
HTML	HyperText Markup Language
IAIDQ	International Association for Information and Data Quality
IBS	International Banking Statistics-
ICA	International Council on Archives
ICE	In Case of Emergency
ICFRE	Indian Council of Forestry Research and Education
ICS	Improvement of Crop Statistics
IdM	Identity management
IEC	International Electro-technical Commission
IEC	Importer and Exporter Code
IIP	Index of Industrial Production
IMD	India Meteorological Department
IMF	International Monetary Fund
IMS	Information Management System
IOTT	Input Output Transaction Table
ISAM	Indexed Sequential Access Method

ISI	International Statistical Institute
ISO	International Organization for Standardization
IT	Information Technology
JDBC	Java Database Connectivity
KNIME	Konstanz Information Miner
LB	Labour Bureau
LDAP	Lightweight Directory Access Protocol
MANOVA	Multivariate Analysis of Variance
MAR	Missing At Random
MCA	Ministry of Corporate Affairs
MCI	Medical Council of India (MCI)
MDDB	Multi Dimensional Database
MDM	Master Data Management
MDR	Metadata Registry
MHRD	Ministry of Human Resource Development
MIS	Management Information System
MLE	Ministry of Labour and Employment
MLP	Major Livestock Products
MOF	Master Office File
MOSPI	Ministry of Statistics and Programme Implementation
MSD	Metadata Structure Definition
MSFTI	Monthly Statistics of Foreign Trade of India
MTPD	Maximum Tolerable Period of Disruption
NABARD	National Bank for Agriculture and Rural Development
NACO	National AIDS Control Organisation
NAS	National Accounts Statistics
NASA	National Academy of Statistical Administration
NASSCOM	National Association of Software and Services Companies

NCTE	National Council for Teacher Education
NDDB	National Dairy Development Board
NDP	Net Domestic Product
NES	National Employment Service
NHB	National Horticultural Board
NHM	National Horticulture Mission
NLTK	Natural Language Toolkit
NSC	National Statistical Commission
NSO	National Statistical Organisation
NSSO	National Sample Survey Office
OASIS	Organization for the Advancement of Structured Information Standards
OCR	Optical character recognition
ODBC	Open Database Connectivity
ODMA	Open Document Management API
OEA	Office of the Economic Adviser
OECD	Organisation for Economic Co-operation and Development
OID	Object Identifiers
OLAM	Online Analytical Management
OLAP	Online Analytical Processing
OMR	Optical mark recognition
OTFE	On-the-fly encryption
PCI	Per Capita Income
PII	Personal identifying information
PIN	Personal Identification Number
PMSD	Planning, Monitoring and Statistics Division
QE	Quick Estimates
QGDP	Quarterly Gross Domestic Product
RBI	Reserve Bank of India

RGI	Registrar General of India
RM	Records Management
ROC	Registrars of Companies
ROI	Return on Investment
RPO	Recovery Point Objective
RS	Remote Sensing
RTO	Recovery Time Objective
SAC	Space Applications Centre
SAS	Statistical Analysis Software
SASA	State Agricultural Statistics Authorities
SDC	Secured Data Center
SDDS	Special Data Dissemination Standards
SDMX	Statistical Data and Metadata eXchange
SFTIC	Statistics of Foreign Trade of India by Countries
SLR	statutory liquidity ratio
SOAP	Simple Object Access Protocol
SPARK	Service and Payroll Administrative Repository
SPSS	Statistical Package for Social Sciences
SQL	Structured Query Language
SRSA	State Remote Sensing Agencies
SSA	Sarva Shiksha Abhiyaan
TCD	Technical Committee of Direction for Improvement of Animal Husbandry Statistics
TMNE	Technology Mission for development of Horticulture in North-East
TRS	Timely Reporting Scheme
UDI	User Data Integration
UIMA	Unstructured Information Management Architecture
UNSD	United Nations Statistical Division
UOM	Unit of Measure

URL	Uniform Resource Locator
UTs	Union Territories
VSAM	Virtual Storage Access Method
WOL	Web Ontology Language
WPI	Wholesale Price Index
WSS	Weekly Statistical Supplement
XMI	XML Metadata Interchange
XML	Extensible Markup Language

Introduction, Executive Summary and Observations & Recommendations

Introduction, Executive Summary and Observations & Recommendations

Introduction:

In India, the Ministry of Statistics and Programme Implementation (MOSPI), various Ministries and Departments of the Union Government and State governments are engaged in the collection, compilation and dissemination of official Statistics. A large number of the data series are generated as a by-product of the general administration of the organs of the Centre and the States/UT based on the records of the concerned offices, as also a product of the administration of particular Acts of the Government and Rules framed there under them. This system generates data on a wide range of subjects. However, neither they are in readily usable form nor up-to-date in public domain.

As official statistics has emerged as a barometer for objectively assessing the quality of governance in the country, An increasing requirement and demand is being felt for up-to-date, reliable and credible data sets (being made) available through single window access. The need for local micro level data in various development oriented programmes/schemes are gaining momentum following the decentralisation process set in motion by the 73rd and 74th Constitutional amendments that gave greater responsibilities and powers to the *panchayats* and nagar palikas. Therefore, the thrust of the recommendations is on improving the existing mechanism of data availability with the trust and responsibility lying with MOSPI- being the nodal agency - for all statistical matters in the country.

Executive Summary and Observations:

"Data Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an organization." Alternatively, the experts define as "Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets."

The challenge for a decentralized data system of the kind currently in place in Government is even more daunting. It should be noted that many of these skills and technologies are widely understood and practiced by private sector in India. The issues are ones of resources and organisation, rather than intrinsic difficulty. Data management encompasses a range of disciplines focusing on managing data as a valuable organizational resource.

Various challenges that are likely to be faced in the data management of official statistics are elaborated in Chapter-1. Issues such as how priorities are set, need to know and right to share, managing a decentralized system, interoperability, need for accessibility and communicability, need for scanning of published data before conversion into defined data format, need to accept data generated by non-government actors, consistency in definition and standards, draw upon insights into problems, vertical and horizontal issues, technological and technical issues, and training intervention/ capacity building are discussed. Each topic gives an illustrative picture of the challenges which is expected to equip the reader to deal and tackle challenges.

The Data Management has now emerged as a vast, diverse and complex field, encompassing technological, administrative, privacy and security dimensions. In order to provide some indication of the potential scope of the field, chapter-2 of this report draw upon the literature to provide an indication of the scope of the discipline. Instituting such a discipline within the confines of even a single large organisation is difficult. The key data management disciplines have been dealt with in a simple framework so that even non-professionals can comprehend easily and adopt the solutions efficiently. Disciplines in data management like Data governance; Data Architecture; Analysis and Design; Database Management; Data Security Management; Data Quality Management; Reference and Master Data Management; Data Warehousing and Business Intelligence Management; Document, Record and Content Management; Meta Data Management; Contact Data Management: SDMX and Secured Data Centers have been dealt with in brief in the Chapter-2 so that the fundamentals of data management get focused and understood. Since this conceptual background is the foundation stone on which data management is to be built, without the knowledge and shared appreciation of these concepts, no

systematic and professional data management can be successful, that too in a country like India where all types of difficulties, heterogeneity, mindset, levels of IT usage and bureaucratic hurdles are likely to be confronted. The entire conceptual framework for data management has been set out in this Chapter-2.

Recommendations:

Finally, Chapter-3 provides some recommendations for consideration of the National Statistical Commission which has constituted the Committee on Data Management. Some of the recommendations can be taken up immediately and some over a medium and long-term horizon keeping in view the migration from current data system to a technology-driven system to manage the data in the Indian statistical system. This would help the system to cater to all types of data needs of the user community say micro, macro, aggregates, derived data set, query based and data mining using Data warehousing technology embedded within an OLAP/OLAM architecture. The recommendations that could be implemented immediately over a short period, say within three to six months , inter-alia, include the following:

1. On every Ministry's web-site, a great deal of (lot of) information, data, reports, circulars, orders etc., are uploaded and forgotten. As a result some are up-to-date and some are obsolete. All data sets are to be identified and converted into portable format.

2. MOSPI, being the nodal agency, should enhance its resources in all angles like latest Hardware, Software and Technical manpower to meet the ever increasing and challenging requirements to build data management capability at national level.

3. MOSPI should transform all its data sets to a portable data format in a unified and harmonized manner.

4. All data sets stored and managed in MOSPI, other line Ministries data sets and data sets from State Governments should be loaded into a main server with all meta data details after the application of ETL tools.

5. There should be a mechanism to update and load the data as and when arrived or generated.

6. Data sets should be split into micro data and macro data sets. For example, Socio-Economic Surveys, Annual Surveys of Industries, etc. micro data is available which could be provided after ensuring that the confidentiality part is not compromised.

7. For simple analyses, Table generation and Report generation suitable and compatible software packages should be integrated into the system to the benefit of wider range of users.

Though MOSPI had been a pioneer in introducing Computer and its application in early sixties, it could not keep pace with the ever and fast changing technology due to various barriers and constraints. MOSPI, the prime source of data, need to assimilate the technological advancements and enhance integration of IT applications and usage in the national statistical system. The section under long run, recommend state-of-the-art IT application and integration of IT solutions in the national official statistical system. Though the recommendations are essential, they are time consuming due to complete overhaul of attitudes, approaches, commitments, harmonization and integration of data and migration from traditional set-up to technology driven set-up. This approach integrates hi-end data warehousing and data mining solutions embedded with online analytical processing (OLAP) and online analytical management(OLAM) architecture for data stream. The following are recommendations which are desired to be implemented within say one or two years:

1. First and foremost requirement is to define and adopt a uniform data format right from grass root level to top most level.

2. After having defined and standardized data set from different ministries, State governments and other data producing agencies, integrate the data set and transform and load them into centrally managed data warehouse server. Here application of ETL tools are necessary to transform the different data sets and load into compatible data warehousing server. Meta data and data-marts would also be taken into account.

3. OLAP and OLAM servers should be built along with the nation wide data warehouse server to enable the users to have analyses, query based filtering mechanisms, generate tables and reports and go into data mining solutions.

To achieve, it is essential to have dedicated team of officers well versed with technology and statistics, cooperation of all the line Ministries and of course dedicated connectivity among MOSPI, line Ministries, State Governments and other data producers a must to succeed in this effort. User would be able to access all types of data say micro, macro and derived data from a single window.

Some of the other recommendations for data collection, data dissemination of confidential data and data sharing are:

4. A sound data collection process based on technology and scientific methodology would yield quality data and subsequently result into credible data at national level on aggregation. To keep this procedure on par with international practices, Computer Aided Telephonic Interview(**CATI)** and Computer Aided Personal Interview (**CAPI)** methods are recommended for data collection. It could be explored in NSS Socio-Economic Surveys and ASI immediately.

5. Set up Secured Data Centers(**SDC**) to access confidential data by the authorized user. The data type could be sensitive like tax data, banking data, criminal records, personal data, etc. All required permission should be taken from the respective data producers to give access to licensed user in a secured environment.

6. Finally, MOSPI should subscribe and take part in Statistical Data and Metadata eXchange (**SDMX**) programme. This helps in harmonisation across different fields of statistics within and outside country. The aim of SDMX is to develop and use more efficient processes for exchange and sharing of statistical data and metadata among national and international organisations and their member countries. The SDMX standards are designed for exchange or sharing of statistical information between two or more partners.

7. To implement the data management project of this mammoth dimension at national level, it may be appropriate to create a new division in CSO with dedicated composition of requisite manpower and other resources (H/W, S/W, Dedicated WAN (Optical Fiber Networks, etc.). As proper coordination with line ministries and State Governments and other national and international agencies are involved, the Data Management Division should be managed by an ISS officer at Spl. DG level with 4 ADG, 12 DDG, 24 JAG, 30 STS/JTS level ISS Officers with suitable supporting Programmers, Database Administrators, System Analysts and other supporting staff at the Centre. In each state the staff size required is 1 DDG, 2 JAG, 4 STS/JTS level ISS officers. This Division could be expanded over a period of time while expanding its functioning.

Chapter – 1

Indian statistical system, challenges in official statistics & data management and remedial measures

CHAPTER-1

Indian statistical system, challenges in official Statistics & data management and remedial measures:

1.1 The Importance of Official Statistics

Given programmatic needs and public interests government agencies share information internally and, as data providers, disseminate extensively to external stakeholders. They make data available to individuals, academics, businesses, and not-for-profit organizations alike. These stakeholders use Public Intelligence for a spectrum of purposes that range from deciding where to live, where to locate a business, or to fueling social activism and political advocacy.

The commoditization of computing power and network access and the rise of the Web have created unprecedented possibilities (and pressures) for participatory, open government, fueled by Public Intelligence. As a result, official statistics have become more important to more people than ever before. That importance only continues to grow, citizen-focused, performance-driven, transparent government.

Open government principles dictate that public administrations must do their best to meet the demand for official statistics, and must disseminate them as accurately, quickly, and usably as they can. The National Statistical Organisation(NSO) must accommodate a very diverse user community with disparate needs as well as an extensive variety of data access and analysis technologies.

Concerns for Official Statistics Providers

The key is opening core government processes to public participation and providing mechanisms that enable and encourage the public to participate.

Achieving openness and participation requires more than good will and a declaration of intent. The trick is balancing principles against practical concerns linked to the production, sharing, and dissemination of official statistics. Relating to official statistics and self-service data access, special, practical concerns include the following:

- Government data must be accurate, authoritative, and timely.
- Data must be statistically valid… and should carry a disclaimer regarding validity conditions.
- Confidentiality must be protected, especially in the face of new technologies that make it easy for external users to join datasets, despite transparency imperatives.
- Interfaces must be accessible to persons with disabilities.
- Data must be distributed in machine-readable formats with sufficient metadata to facilitate the data's use.
- Provision of "mashable" data that can be easily linked to or combined with disparate, other data is a goal, yet the cost to users, who have paid for the collection and production of the data, must be minimal or, in many cases, nil.

- Governments must accommodate secondary users and data aggregators who redistribute public data via value-added services and applications.

Dissemination of Statistics and Technology

Dissemination of official statistics is in its third generation. The first was distribution of statistical tables on paper, in reports and bound volumes. Distribution in printed form limits the audience and does not facilitate secondary analysis.

The second generation, spanning the 1970s through the mid-2000s, involved data dissemination – by both governments and secondary data providers – on magnetic tapes, diskettes, and disks and via early Website query and download interfaces. Only technically oriented users would undertake any form of substantial data analysis, typically using statistical analysis, business intelligence, and spreadsheet software.

The advent of trend toward open government has pushed dissemination of official statistics into a third generation, where access is primarily online, types of use and users are hugely varied, and new access standards and methods are required to support highly diverse applications.

Data Users

The simplest way to describe them is to consider them as ranging from institutions to individuals with secondary providers – commercial data aggregators, university research centers, data archives, and portals – thrown into the mix as multi-role consumer-providers.

Individuals, including most casual users, have traditionally looked for particular, focused data elements: perhaps unemployment statistics over time or a comprehensive statistical profile of their home town. Their data consumption patterns have tended to become much more dynamic, less easily satisfied with pre-built data exploration and display interfaces, instead requiring new abilities to construct custom, hybrid data analysis objects. Their statistical literacy is often lacking.

Institutional users – government internal users, businesses, researchers – have a level of subject-matter and IT expertise higher than that of individual users of official statistics. Their work is project driven, often involving policy and evaluation, supported by commercial data-analysis applications and enterprise-grade IT infrastructure including, increasingly, cloud and software as-a-service computing.

Modern access methods

While many data consumers, including data aggregators and resellers and hard-core researchers, will continue to use older-style interfaces and dataset download options, the prevailing trend is toward data access via application programming interfaces (APIs), Web services, etc.

Agencies are expected to variously support or at least facilitate official portals; de facto portals including Google and other search engines, which have loaded government data and deliver statistics in response to data queries; and global, non-commercial initiatives, notably the emerging "web of data" that has evolved.

Support for these access methods entails preparation and release not only of datasets in computer-processable format but also of metadata sufficient to support stand-off identification, extraction, and use of desired information.

Best practices of data access and analysis

Software tools enabled online services, launched by some of the advanced countries like US, UK, Australia, etc. allow users to dynamically construct ad hoc tabulations, graphs, and maps from primary data sets. Comparable capabilities are very rarely offered to end users by most other government statistical agencies worldwide, which disseminate only prepackaged and highly aggregated data. These countries provide access to data in three ways: via a "raw" data catalog linking to download of machine readable, platform-independent datasets; a tools catalog with links to agency tools or Web pages; and through a proper catalog with links for geographic information. The initiative is supported by an effort to standardize metadata, formats, and dissemination policies.

The official statistics common ground

The primary objective in sharing government data with a wide spectrum of stakeholders is the realization that the official data is a core asset that enables participatory and open governance. Achieving this goal is an on-going process. The process involves a balancing act. It aims at facilitating open access and accommodating the widest variety of users, applications, and access methods while attending to official-statistics provider concerns relating to accuracy, timeliness, confidentiality, accessibility and other special needs.

1.2 Current scenario of National Statistical System in India:

As the largest functioning democracy with varying culture every two hundred km and multitude of spoken languages makes the data management system highly complex. As a consequence of this plurality data needs are very varied and highly demanding. The Ministry of Statistics and Programme Implementation (MoS&PI) is the nodal agency for the planned and organized (a) development of the statistical system in the country and coordination of statistical activities among statistical agencies in the Government of India, State Governments as well as meeting requirements of the International Agencies like UN, World Bank, IMF, OECD, etc. It has two major organizations, namely, Central Statistics Office and National Sample Surveys Office(NSSO) combined together known as National Statistical Organisation (NSO). The collection of statistics on different subject areas, like agriculture, labour, employment, trade, industry, etc. vests with the designated administrative Ministries.

Generally, the statistical information is collected as a by-product of administration or for monitoring the progress of specific programmes/schemes.

Large-scale statistical operations like the Population Census, Annual Survey of Industries, National Sample Surveys (Socio-Economic Survey by NSSO), Economic Census (by CSO), Agriculture Census, Livestock Census, etc. are generally centralised, and these cater to the needs of other ministries and departments, as well as State Governments. Similar set up exists in the States. At the apex level in the States is the Directorate of Economics and Statistics (DES), which is formally responsible for the coordination of statistical activities in the State. They bring out statistical abstracts and handbooks of the States, annual economic reviews or surveys, district statistical abstracts, and State budget analysis; work out the estimates of the State Domestic Product and Retail Price Index Numbers and engage in such other statistical activities as is relevant to the State.

Large amount of data exist(s) at the Centre as well as in the States. But they are not integrated either horizontally or vertically. Further getting timely, reliable and credible data on many subject fields are too difficult and time consuming. And in many cases data are not made available at all due to various factors like late processing, dissemination, etc.. Except few statistics namely National Accounts Statistics(NAS), Index of Industrial Production(IIP), Consumer Price Indices(CPI), Wholesale Price Indices(WPI), NSS Socio-Economic results and select Agriculture Data, there is no clear advance release calendar available. India is yet to adopt the best practices available in the world to collect, compile, process and disseminate the data. Therefore, to achieve the above objective, the Data Management Committee has been formed under the aegis of National Statistical Commission with the objective of bringing the operations of national statistical system and its dissemination methods to the level of best practices illustrated above. Before going into the details of operationalization of proposed data management, it is pertinent to discuss various issues involved in data collection, data processing and dissemination and the mandate and obligation of NSO in MOSPI.

1.3 Need to know and right to share,

Every citizen likes to have in his possession timely, reliable, relevant and quality data and information at his door-step. An effective combination of analyzed and visualized data help to make the appropriate decision when one needs to know more about national or regional statistics. This also helps to discover patterns and insight to provide the basis for better understanding and decision making with the same tools and methodologies that today are used by many national and regional statistical agencies. Passing on knowledge from analysis and storytelling through website or blog helps in easy comprehension of data.

Official Statistics are about describing the situation in a country regarding the economic, social, cultural and political conditions and the development thereof over the years. This means that before we can speak about any number we need to know what these numbers are about. That is the subject matter (content) aspect of

statistics. This type of understanding can be brought about in a simple way. This means that we can address these conditions based on a common sense approach. It also can be done in a very sophisticated way. In that case, the statistical experts need to be informed about all the relevant social and economical theories that are of importance. Learning about statistics can be done in a simple way and in a very sophisticated way. In reality combinations of these two approaches are always needed, in different degrees. The subject matter approaches are needed to design the surveys, more specifically to design the questionnaires and the frameworks for the analyses of the results.

The second aspect of official statistics is the methods. Some think that it is the first aspect of official statistics but that is not so. None can use any method when one does not know what they are about. When one knows what the topic is about one can select and apply the methods that are relevant. The concept of a method can be defined in different ways. It can be defined in a very limited way. That is for instance when we define what formula should be applied. The other way of looking at a method is to see a method as part of a much larger and much more complicated process. This is called the statistical production process and it also may be called the statistical value chain. When we look at statistics in this way we realize that the total of activities that make a statistical organization functions well is part of the "overall method of production of statistics" which is applied. This approach includes the institutional tools (methods) and the management tools (methods) that are used.

A statistician who wants to apply his formula to solve a specific mathematical problem in the end also will depend whether all the other parts of the organization are functioning well enough. These other parts also need to function sufficiently in order to make sure that this statistician can apply his formula. In this "overall" concept of the method even the bookkeepers of the accounts of the organization are part of the methods that are applied. When the statistician is not paid, he will not apply his formula.

The third aspect of statistics is about the results. Results need to be sound enough to be published. What is sound can be defined. This is part of the quality assurance approach in statistics. Within the context of producing results we need to pay explicit attention to the quality aspects of these results. This is about how these results are disseminated and made accessible to the public. Here the concept of meta-data is relevant.

1.4 Issues of decentralized system,

An important issue currently in India involving the production and dissemination of statistics is the growing role of statistical departments of sub-national administrative

units (states and local bodies) in data production. Increasingly, autonomous state statistical departments are conducting surveys on a variety of subjects that, in most cases, are already surveyed by the NSO. Since these departments function in the overall federal structure of governance with budgetary autonomy, the Central Government has limited means to interfere in their statistical activities. Many of these surveys utilize different methodologies from the ones used by the NSO and, as a consequence, produce different results, causing incomparability and inconsistency in inference at the users end.

On the one hand, data for policies of direct economic intervention and control are needed less than before. But on the other hand, more refined and precise information, based on more comprehensive and accurate statistical data, is now required for polices of indirect economic stimulation by the government, as well as for decision making in the private sector in today's complex economy.

With limited resources made available for sample surveys, it has become more difficult to extend or expand existing statistical surveys, or to introduce new ones. This has led to the recognition that rather than relying solely on statistical surveys in compiling statistics, existing administrative records and registers must be more widely and efficiently utilized.

At present MOSPI is the nodal agency for coordination, but its powers are insufficient for the task, and empowerment of MOSPI as a stronger agency for the coordination and total planning of official statistics is called for, may be through the Collection of Statistical Act 2008. We identify some key processes that would need to be addressed within the normal organisational and administrative frameworks that exist in the country to enable devolved or decentralised systems to be effective.

The first requirement is the need to nurture a harmonised culture of statistical professionalism, promoting practices of evaluation and quality assurance and emphasising the use of best practices. Without such a shared approach the problems of decentralised or devolved systems may prove to be insurmountable.

Second the system needs to be based upon mutual benefit to the parts insofar as this is achievable. The essential role that each part plays needs to be recognised and valued. The strengths, and in particular the professional skills and expertise that exists need to be recognised and these should be reflected in the division of roles and responsibilities so as to achieve the best outcome.

Third there is a need for some system of user interface that is open and transparent and that tries to response to the needs of users (and in the context of devolved authorities) at all levels. No statistical system can respond to all needs but a process that exposes the competing needs and attempts to establish the 'public interest' in some sense in meeting them provides the basis for resolving conflicting priorities.

Fourth there is a need for a co-ordinated planning and priority setting system that spans the key parts of the system and which builds on the user consultation processes. Inevitably there will be tensions and conflicts of priority and there needs to be a process for conflict resolution although how these are resolved will vary with

national cultures. Where funding is provided to different parts of a decentralised or devolved system separately then the process needs to be able to identify where inadequate funding to one part of the system is impairing the effectiveness of the whole.

Fifth, it is essential that common concepts, classifications and standards are used so that statistical estimates are comparable and that statistics derived from different sources may be combined as needed. Some accepted process for establishing these decisions is essential including resolving conflicts that will occur.

Sixth there is a need for processes of quality evaluation and quality assurance that span the system as a whole and which are conducted in the spirit of quality improvement rather than apportioning blame. The user consultation process can support quality assessment since often, knowledgeable users will have an appreciation of gaps and inconsistencies and of the overall quality achieved.

Finally, it is important that the statistical system is receptive to international developments, in terms of changing concepts, the need for new methods, standards and classifications. It needs to be responsive also to the changing needs of supranational and international agencies. The processes of user consultation, priority setting and quality assurance and evaluation must inform this involvement with the international statistical community and this implies that there is a need for strong co-ordination within the national system.

1.5 Vertical and horizontal issues,

The main features of the Indian Statistical System can be summarised as:

The Administrative Statistics System is its major component; It is laterally and vertically decentralised. In it, not only data collection but also compilation, processing and preparation of results are carried out by the States for most of the sectors. It is the State-wise results, which flow to the Centre, and statistics at the all-India level are obtained, more often, as the aggregates of State-level statistics.

Subjects such as money and finance, international trade, balance of payments, incorporated businesses, have meaning only at the all-India level. There are sectors, which straddle across more than one State and for which statistics are collected directly by Central agencies such as railways, postal services and telecommunications. Statistics on both types of subjects are collected by the Central Administrative Statistical System. All other Administrative Statistics are collected by the State Statistical System. Administrative statistics provide information that is relevant to the working of the Departments. They serve the major purpose of aiding the Departments in the execution of their administrative functions of implementation and execution of different Acts, Rules and Regulations with which the Departments are entrusted. Consequently, the concerned Departments have a vital interest in the proper collection of the administrative statistics.

There are other advantages too in the system of collection of statistics through the administrative set up. The collection of data by departmental agencies does not involve special costs. The collection is oriented to definite purposes, and the record and verification of information is part of administration. Departmental agencies and officials have not only good knowledge of the subject, but also of local language and local conditions, especially rural. Information collected is relevant and direct, and the respondents do not have to make calculations before answering a query. It is handled by agencies that have special knowledge of the subject. Finally, there is an identifiable purpose in their data collection and they are in the best position to interpret the data. All this has lent a solid foundation to the decentralised administrative statistical system, and in turn, to the Indian Statistical System. An impression is carried by many that data collected by substantive Government departments are likely to suffer from bias. Therefore, they suggest that an independent agency should collect data to ensure *objectivity*. But, ignorance should not pass off as objectivity, making the solution worse than the problem. While the impression might be true for certain departments at certain times, it is easy to overstress the point as a justification for the solution suggested.

A variant of this system is the one where a central agency collects data directly from the district offices of the State Government departments. When the State departments have to process the data and produce results, before they transmit them to the Centre, they are *per force* required to pay attention to timeliness in collection of data and their quality, and take corrective actions. But when the reporting units are to send the data directly to the Centre, which is a far distant agency, timeliness and quality of data will be affected adversely. For the States will have no responsibility for either. Worse, the State-level results, in this scheme of things, will become disaggregations of the National totals, and the States will be dependent on the Centre for State-level statistics, which are in fact in their domain.

But over the years, the Administrative Statistical System has been deteriorating and 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 few sectors: agriculture, labour, industry, etc. 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.

A similar problem is that of the weakness of lateral coordination, which has come to be viewed as another major problem. The CSO carried out its function of coordination mainly by means of the technical committees or working groups, either appointed by it, generally under the Chairmanship of the CSO Director General, or by the Ministries in which case the CSO was generally represented on them. The other mode was the bi-annual Conference of the Central and State Statistical Organisations (COCSSO) organised by the CSO. The COCSSO provided a forum for exchange of views and experiences concerning development of statistical activities in the country.

The experience in coordinating with the Ministries by CSO has not been satisfactory. The available institutional arrangements for coordination in the shape of Technical Working Groups on various subjects or Committees were either not continued or were terminated by the Ministry of Statistics. The strengthened coordination with various data producing agencies is an important tenet of an efficient data management system. National Statistical Commission(NSC) set up in 2007 attempts to address these issues.

1.6 Priority

To deliver tangible results within an acceptable time frame, NSO must prioritize categories of data to be brought out at regular intervals with pre formatted advance release calendar so that the users are aware of the availability of statistics.

Experts advise that organisations take a risk management approach to prioritizing data release. "There are different types of information assets in the organization with different levels of value and different levels of exposure." Some are to be released to public as soon as the results are found/estimated.

National Statistical Organization(NSO) may prioritize core statistics over other statistics. Even within core statistics, NSO may prioritize various sectoral data for policy purposes. National accounts statistics, Agriculture production statistics, labour statistics, education statistics, price statistics, health statistics, etc. are key statistics on which data should be disseminated with minimum time lag and maximum disaggregated level including that of micro level data. NSC has been seized of finalizing the core statistics and data within core statistics should be given priority which caters to the planners and policy makers.

1.7 Interoperability,

Interoperability is a property referring to the ability of diverse systems and organizations to work together (inter-operate). The term is often used in a technical systems engineering sense, or alternatively in a broad sense, taking into account social, political, and organizational factors that impact system to system performance.

The experts define interoperability as the ability of two or more systems or components to exchange information and to use the information that has been exchanged. While interoperability was initially defined for IT systems or services and only allows for information to be exchanged, a more generic definition could be this one :

Interoperability is a property of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, without any restricted access or implementation.

This generalized definition can then be used on any system, not only information technology system. It defines several criteria that can be used to discriminate between systems that are "really" inter-operable and systems that are sold as such but are not because they don't respect one of the aforementioned criteria, namely :

- non-disclosure of one or several interfaces
- implementation or access restriction built in the product/system/service

Opening public information has recently become a trend in many countries around the world and India should not be far behind. Online government data catalogues with national, regional or local scope act as one-stop data portals providing descriptions of available government datasets. These catalogues though remain isolated. Potential benefits from federating geographically overlapping or thematically complementary catalogues are not realized. Uniform format or schema should be in place as an interchange format among data catalogues and as a way of bringing them into the Web of Linked Data, where they can enjoy interoperability among themselves and with other deployed datasets.

At present, most of the data released by NSO are in PDF format. They are unfit for further analyses and requires enormous amount of time to reformat them in a portable data format. NSO must aim in publishing all its data, preferably, using data warehousing and data mining tools.

1.8 Need for accessibility and communicability,

Statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.
The MOSPI should ensure that the results it produces are presented clearly and in an orderly manner, whether they are disseminated on paper or via the internet. They are in public domain and are easily accessible to anyone.
Presently, the MOSPI disseminates its products and publications accompanied by press releases to facilitate understanding of the data being published and avoid errors of interpretation. Subject specialists also answer questions from journalists wishing to obtain detailed information about how the figures are produced.

Information accessibility means the ease with which all categories of users can know that the information exists, find it, and import it into their work environment.
The objective of the MOSPI in terms of dissemination is to do all it can to promote the use of the information it produces. The MOSPI produces statistical results of national interest as stipulated in its allocation of business rules, and makes them available free of charge on the internet. It supplies its publications free of charge to those media which promote them in their materials.
The publications catalogue and others reports are displayed on the website. The services and search tools to help users find the data they need are essential to accessibility. The MOSPI answers anyone who asks a question by e-mail, telephone or letter to try to help them use the Ministry's web-site (www.mospi.nic.in). It takes very regular action to improve quality in the way it processes the requests sent in.

The MOSPI also offers web users an integrated Data Where housing tools giving access to the official statistics published for analytical purposes. The data warehousing portal offers users practical and harmonized access to many rounds of socio-economic surveys data at micro level.

The dissemination of detailed data, whether derived from surveys or administrative source files, is organized within the limits to meet the particular needs of statistics specialists. Individual data on persons or households is made available on website. These files are presented in anonymous form preventing any identification. Finally, access to individual data supplied for the purposes of official statistics or scientific or historical research. The process of dissemination of micro level data on socio-economic statistics using data warehousing and data mining technology should be expanded to cover other data sets.

1.9 Need for scanning of published data before conversion into defined data format

NSO generates a large volume of data. Most of them are disseminated in a publication form and weeding them out periodically is not only desirable but inevitable due to shortage of office space. Scanning and storing of these published data in a digital form is a must before being weeded out. MOSPI is digitising most of the Administrative files and digitising of periodicals and publications are also required.

Document scans are often processed using OCR technology to create editable and searchable files. Most scanners use appropriate device drivers to scan documents into TIFF format so that the scanned pages can be fed into a document management system that will handle the archiving and retrieval of the scanned pages. While paper feeding and scanning can be done automatically and quickly, preparation and indexing are necessary and require much work by humans. Preparation involves manually inspecting the papers to be scanned and making sure that they are in order, unfolded, without staples or anything else that might jam the scanner.

Indexing involves associating relevant keywords to files so that they can be retrieved by content. This process can sometimes be automated to some extent, but it often requires manual labour performed by data-entry clerks/operators. One common practice is the use of barcode-recognition technology: during preparation, barcode sheets with folder names or index information are inserted into the document files, folders, and document groups. Using automatic batch scanning, the documents are saved into appropriate folders, and an index is created for integration into document-management systems.

A specialized form of document scanning is book scanning. Technical difficulties arise from the books usually being bound and sometimes fragile and irreplaceable, but some manufacturers have developed specialized machinery to deal with this.

Some organisations have made a transition to a paperless office. The way they can do this is by having the ability to scan any sized paper that comes into the office. If the organisation/Ministry doesn't want to take on this responsibility there are document scanning services available. **For achieving saving in finance, office space, manpower resources, this service could be made use of.** The ability to have multiple ways to store files is a long-term, and hazard protection safety net. One will be able to keep files for years, or decades in a small media format. One can have it copied on the server, on a disc, and stored at another secure location. With this duplication ability, the files should be safe from most natural disasters.

When the paper is removed from the office the storage areas become usable space. If one store paper in boxes in a separate room, once all is removed, one has an extra office. The staff needed to manage, file, and organize paper might be eliminated or moved. All of the costs for office supplies that are needed to support paper filing now become savings for the organisation/Ministry. No file cabinet, no hanging folders, no manila files, and no labels.

One can sell the clients/users on the speed and ease of how you can transmit needed information. Articles, documents, and forms are easily emailed or faxed from digital storage. The turnaround time for contracts and correspondence can easily be reduced.

Even if one has secure information to be scanned, (don't worry) scanning is available at the desired location. The procedure is readily accessible to monitoring and certification can be given for complete discretion. General office files can be done on-site or off-site. The files can be returned in the paper form or certified shredding might be offered.

1.10 Non-official data providers

The generation of data by the Government is primarily based on its needs for decision support and is prioritized to inform the citizens on varied affairs in the country. This requirement, in many cases is stable and systematized over time. However, the data often is required on many other aspects to meet specific and occasional requirements. The constraint of flexibility within the government to mobilize resources for such data needs often necessitates the sourcing the data from non official sources. However, there is no guarantee that data collection by private agencies would not be subject to its own biases, that monitoring the quality of such data would be difficult, and finally, that the competence of such agencies have to be established.

Data collection by private agencies is complex though important and it is difficult to find any agency in India which can provide data on national level on any parameter on regular basis. RGI, NSS, some central Ministries have mechanism to provide data on regular interval for select parameters.

It needs to be appreciated that collection of primary statistics needs well trained human resources. In view of the constraints in the government to create new posts

by considerations of economy in long-term costs and of obviating the problems of staff management, and one cannot expect too much resources, in terms manpower and money, being made available to the data collection activities. Consequently, capabilities available with non-governmental sector could be made use of for data collection to complement official data collection agencies. If an established private sector organization provides data at national or sub-national level where an official mechanism is not able to meet the requirement it may sometimes be inevitable to resort to outsourcing to non-official data providers, with a suitable mechanism of ensuring reliability, credibility and soundness of such data.

1.11 Consistency in definition and standards,

Consistency refers to logical and numerical coherence. Consistency over time, within datasets and across datasets (often referred to as inter-sectoral consistency) are major aspects of consistency. In each, consistency in a looser sense carries the notion of "at least reconcilable." For example, if two series purporting to cover the same phenomena differ, the differences in time of recording, valuation, and coverage should be identified so that the series can be reconciled.

Inconsistency over time refers to revisions that lead to breaks in series stemming from, for example, changes in concepts, definitions, and methodology. Inconsistency within datasets may exist, for example, when two sides of an implied balancing statement-assets and liabilities or inflows and outflows - do not balance. Inconsistency across datasets may exist when, for example, exports and imports in the national accounts do not reconcile with exports and imports within the balance or payments.

1.12 Draw upon insights into problems,

(a) Use of sound methodology

Although technically sound science is not necessarily ethical science, the failure to use sound technical methods can be so flagrant or long-standing as to present serious ethical issues. In official statistics, this sort of situation may arise when a seriously flawed or outdated methodology continues to be used by an agency long after its shortcomings have been identified and alternative approaches explored, but for inappropriate reasons (for example, political pressures or a false sense of institutional pride), the flawed approach remains in use, sometimes for decades. It should be understood that not every error or flawed procedure is an indicator of an ethical lapse. Official statistics is carried out in a world of deadlines and limited resources. In some cases, the methodological issues are not been paid desired attention. Indeed, a sound statistical system would need to upgrade itself in capabilities, orient to new requirements, benefit from examining the criticism and deficiencies., explore ways of overcoming them, and introducing needed improvements. But in the end there is both an ethical and scientific imperative to make the needed improvements. Not to do so, at some point, can become both a scientific and ethical failure. In India, sampling methods for ASI and NSSO-Socio-

Economic surveys are some of the examples where the procedures of sampling has been refined and customised to meet the subject requirements and improving the data quality.

(b) Protection of confidentiality

The roots of the concept of statistical confidentiality and the protection from harm attributable to cooperating with statistical inquiries can be traced back to the Hippocratic oath were physicians agree not to cause harm to their patients and not to gossip about information obtained in the course of their professional work. The modern concept of statistical confidentiality evolved as a means of encouraging businesses/households to report accurately by assuring them that business rivals, muckraking journalists and Tax agencies would not have access to the information they provided, except as statistical aggregates. However, legal protections relating to statistical confidentiality in India only pertain to identifiable microdata. Unfortunately, direct harms have arisen through the use of mesodata (tabulated data for very small geographic units, such as blocks or enumeration districts) to operationally assist in targeting vulnerable population subgroups for internment or worse. Given this experience, and in the absence of clear legal protections relating to mesodata, statistical agencies, together with their leadership and staff, are under heavy ethical obligations to provide as wide a protective net as possible over mesodata pertaining to such vulnerable populations.

c. Integrity of the statistical agencies and the national statistical system.

An important focus of the fundamental Principles of Official Statistics is the maintenance and enhancement of the integrity of the national statistical system. Threats to integrity can arise in a number of ways, including, among others, arbitrary manipulation of concepts, definitions, and the extent and timing of the release of data, doctoring the actual data released, etc. Certainly, the public at large, data users, and political leaders in and out of the government have a long-term interest in resisting such threats. For agency leadership and staff, individual statisticians and national statistical societies, NSC and the International Statistical Institute (ISI) that interest in resisting such threats becomes an ethical responsibility.

How may one deal with the ethical problems that arise in official statistics?

During the course of work in official statistics one may encounter or anticipate a range of ethical challenges. The issue then arises over what to do.

(a) Coping strategies: A number of options are available in dealing with what one perceives as an ethical problem in government statistical work. The basic elements of an initial response are: speaking up about the perceived problem, consulting with other colleagues informally, establishing a written record, and explicitly informing one's supervisor about these concerns. When documenting these concerns in writing it is useful to indicate how the proposed action or existing practice violates established professional, agency, or government-wide norms for statistical work. In thinking about what to do at any stage, it is important to remember the principle of proportionality of response to threat or harm. Minor harms do not usually justify major action and we should also avoid escalating differences over scientific methods into an ethical controversy.

(b) Prevention strategies: A robust set of prevention strategies is perhaps the best means of avoiding serious ethical problems and of ensuring that any that do arise can be dealt with responsibly and expeditiously. Possible preventative actions include: (1) studying and documenting previous problems; (2) developing and disseminating case studies that illustrate ways of addressing ethical issues based on real or hypothetical examples; (3) developing enhanced models of disclosure risk that take into account both the probabilities of disclosure and the possible harms that may arise from such disclosures; (4) providing education and training on ethics in university and statistical agency training programs, including those specifically tailored for mathematical statisticians and computer methods staff who often have central responsibility for work on disclosure safeguards; (5) developing agency-specific plans for fostering discussions of agency ethical issues and agency specific mechanisms for responding to ethical concerns; (6) developing, with external input and public comment, statements articulating the ethical standards that agency staff and management are expected to follow; and (7) further developing and applying a range of substantive, methodological, operational, and legal safeguards. Statistical agencies that are actively pursuing such preventive strategies will also have a management and staff alert to the ethical issues that arise in their work. Such a sense of ethical alertness is perhaps the single best defence against major ethical tragedies.

1.13 Technological and technical issues

Current strategy for information technology supports our general strategy by stating that information technology in Statistics shall contribute to improved efficiency and development of new possibilities within:

- Data collection (both based on registers and forms)
- Revision of data and production of statistics
- Analyses
- Availability and dissemination of statistics and analytical results
- Office administration and support.

MOSPI plans for introducing the use of business registers as a basis for enterprise statistics. (BUSSINESS REGISTER WILL PROVIDE THE FRAME, THAT WOULD FECILITATE FURTHER DETAILED SURVEYS). Business Register envisages to strengthen the data management process through systematised records, making it comprehensive for further compilation of data using alternative methods. Use of registers requires computer systems of large capacity in general, and possibilities for mass storage in particular. Data which has to be collected directly from respondents should be collected and transferred effectively with a minimum of work for the data suppliers. Electronic Data Interchange (EDI) and optical reading of forms are relevant techniques in this context. Processing of data within MOSPI (revision, aggregation, analyses and presentation) and eventually dissemination also puts different requirements on computer systems and software.

The IT strategy outlined from the requirements above has been summarised as follows:

- Migration to Client-Server technology from the stand alone system

- Selection of standard software and implementation throughout the country.
- Development of reference databases for official statistics and metadata
- Data collection and dissemination via using IT
  i. Use of CAPI, CATI etc for data collection through a standardised format
  ii. Use hi-end Fiber Optics enabled networking for data transmission from field to higher level (District, State, etc.)
  iii. State of the art software for data processing at various level
  iv. Dissemination of micro, macro, derived statistics through dedicated web-portal and implementation of data warehousing and data mining technologies both at national level and at State level.

One should realise that the variety and different nature of the tasks of a statistical institution require different technical solutions, and achievements and success of technological change vary for the different tasks. The difficulties and success of the described technological change have also varied over the years and in the organisation. There are a few examples of introducing immediately electronic data collection (such as the computer aided interviewing), but it is in particular within dissemination of statistics and office administration that technological change most easily would bring positive results. Development has been slow with regard to the migration of large statistics production systems from the PC to new technological platforms (SAS, ORACLE, SPSS, etc.), where we have to meet considerable obstacles in treatment of large amounts of data resulting from linking administrative Ministries. Management of information technology in a statistical institution is a difficult and big challenge because:

- Technology is new, and develops rapidly
- Technological experts tend to put ambitions too high
- New technology implies and is dependent on new organisation of work processes
- Organisation of technology is difficult, and poorly defined responsibilities and links to top management may be a problem
- Problems are often met with more resources; it is hard to admit that lack of IT experience may be the issue.

Development in and with the help of information technology normally requires extra input from experienced personnel. However, the tasks of a NSO require continuous production and dissemination of data at regular interval. This makes it necessary to have increased resources for the development period, and costs will be higher in the short run even if the goal is increased efficiency in the long run. That is why development and change of technology is a difficult and time consuming process.

Most of the administrative data systems which are important for the production of statistics can be linked to the business registers by identification numbers assigned here. Information technology is the backbone of the activity of a national statistical organisation, but managing it is difficult due to several reasons of which rapid changes, the dependence of specialists and organisational issues are important. It is a general experience that technological changes require more time and resources than foreseen. Application of IT in Statistics to succeed, one should understand the following:

- Ambitions should not be put too high.

• Decisions on technological change must be made clear by top management, and followed by information and necessary resources for the implementation.
• Strong co-ordination, well-functioning co-ordinating bodies and project organisation across a decentralised organisation is necessary.
• New technology implemented and applied should have been tried out in other institutions first.
• Some use of external consultants is convenient, but this requires corresponding internal resources to ensure follow-up of results.
• The issue of human resources is crucial, and consolidation of staff and experience calls for flexibility.

Success story in integration of IT in statistics is (i) Processing of 5-th Economic Census data using OMR/ICR technology, (ii) migration of compilation of national accounts statistics from manual to computerized process, (iii) linking of all State Income Groups of all the 35 States and UTs through GDPNET, a Wide Area Network (iv) Computerization of processing of NSSO Socio-Economic Surveys data and release of results quickly and reducing the time lag considerably (v) Computerization of processing of ASI data and reducing the time lag in release of results, (vii) Online transmission of price data for CPI(New Series) through web Portal and (VIII) Implementation of Data warehousing solutions for socio-economic surveys data.

1.14 Training intervention/ capacity building

Capacity building often refers to any assistance that is provided to entities which have a need to develop a certain skill or competence, or for general upgrading of performance ability. **Capacity Building is much more than training** and includes the following:

- Human resource development, the process of equipping individuals with the understanding, skills and access to information, knowledge and training that enables them to perform effectively.
- Organizational development, the elaboration of management structures, processes and procedures, not only within organizations but also the management of relationships between the different organizations and sectors (public, private and community).
- Institutional and legal framework development, making legal and regulatory changes to enable organizations, institutions and agencies at all levels and in all sectors to enhance their capacities

Capacity building is defined as the "process of developing and strengthening the skills, instincts, abilities, processes and resources that organizations and communities need to survive, adapt, and thrive in the fast-changing world." For organizations, capacity building may relate to almost any aspect of its work: improved governance, leadership, mission and strategy, administration (including human resources, financial management, and legal matters), program development and implementation, fundraising and income generation, diversity, partnerships and collaboration, evaluation, advocacy and policy change, marketing, positioning, planning, etc. For individuals, capacity building may relate to leadership

development, advocacy skills, training/speaking abilities, technical skills, organizing skills, and other areas of personal and professional development. Capacity building is the elements that give fluidity, flexibility and functionality of a program/organization to adapt to changing needs of the population that is served.

Training can be described as "the acquisition of skills, concepts or attitudes that result in improved performance within the job environment". Training analysis looks at each aspect of an operational domain so that the initial skills, concepts and attitudes of the human elements of a system can be effectively identified and appropriate training can be specified.

Training analysis as a process often covers:

- Review of current training
- Task analysis (of new or modified system)
- Identification of training gap
- Statement of training requirement
- Assessment of training options
- Cost benefit analysis of training options

A **"training needs assessment"**, or "training needs analysis", is the systematic method of determining if a training need exists and if it does, what training is required to fill the gap between the standard and the actual performance of the employee. Therefore, training needs analysis is - Systematic method of determining performance discrepancies Causes of performance discrepancies
Training needs analysis includes:

- **ORGANIZATIONAL ANALYSIS –** It includes the analysis of
    - Mission & strategies of organization
    - The resources and their allocation
    - Internal environment- attitudes of people
- **OPERATIONAL ANALYSIS**
    - Determine KSAs required for standard performance
    - Job analysis
- **PERSON ANALYSIS**

- Specific areas of training required by the individual
- Whether an individual is capable of being trained
- The data regarding the person analysis can be collected through-
- Performance data
- Behavioral and aptitude tests &
- Performance appraisal

Performance appraisal can significantly help in identifying the training needs of the employees. Performance appraisal helps to reveal the differences and discrepancies in the desired and the actual **performance of the employees**. The causes of the discrepancies are also found whether they are due to the lack of adequate training or not. The employee can also tell about his training requirements (if any) in his self

appraisal. A performance appraisal after the **training program** can also help in judging the effectiveness of the program.
National Academy of Statistical Administration (NASA) under the Ministry of Statistics and Programme Implementation assess the various requirements of the statistical personnel in the country and imparts training on various subjects ranging from one week to six weeks courses in the area of Statistics, Economics, IT, Management, Administration, Time Series Analyses, Remote Sensing, Natural Resources Accounting, etc.
The established institutions with prime focus on regular production of statistics using sound methodologies such as NSSO have training and capacity building process integrated in their operations.

1.15 Technology Verses Applications:

Advancement in Technology is swift and fast. It is difficult to adapt each and every change in the technology to various day-to-day works. But use of upfront technology can not be undermined. In the area of data management, technology has grown rapidly and data management more simpler than what it used to be. From simple formatting and table generation to data mining and data warehousing, technology plays important role. But when it come to application, many constraints are experienced. Availability of resources, trained manpower, adoptability to the technology and uninterrupted delivery mechanism are to be taken into account. There are a few examples of introducing immediately electronic data collection (such as the computer aided interviewing), but it is in particular within dissemination of statistics and office administration that technological change most easily would bring positive results. Development has been slow with regard to the migration of large statistics production systems from the PC to new technological platforms (SAS, ORACLE, SPSS, etc.), where we have to meet considerable obstacles in treatment of large amounts of data resulting from linking administrative Ministries. Management of information technology in a statistical institution is a difficult and big challenge because:
• Technology is new, and develops rapidly
• Technological experts tend to put ambitions too high
• Organisation of technology is difficult, and poorly defined responsibilities and links to top management may be a problem

Therefore, a reasonable mix of technology and application is a better strategy.

1.16 Legal Framework for providing data

Governments in all its forms, Central, State and Local bodies collect, process and disseminate a large volume of data by spending enormous amount but there is no legal framework to get data by the general public and tax payers. It is left to the will and magnanimity of the government. There should be an appropriate steps in Government of India through suitable legal frameworks to provide its data on demand by the user community.

1.17 Setting up of Data Management Division

NSO generates voluminous data every year. Some are published and some are not even processed for want of resources. If processed, they are abnormally delayed. Being decentralised system, the scenario remains the same at state level as well. The advent of open government system in modern times has pushed dissemination of official statistics into a technology driven, where access is primarily online, types of use and users are hugely varied, and new access standards and methods are required to support highly diverse applications. Public demands have multiplied and requirement of data has changed from simple micro level data to processed analytical data requirement. Wide range of user community faces great difficulty in converting the raw data into their system and platform compatible data set. This needs lot of time and resources at their end. Portable data format is platform independent and is ready to use in any platform or system for further analyses and processes. Further, data integration and matching are huge task at national and State level. These are big challenges to the NSO and NSO should equip itself to meet these challenges and prepared to handle the Data Management and dissemination issues at par with the best practices in the world. To meet the above challenges, NSO must create a full-fledged Division in CSO to look after the Data Management with state of the art facilities in terms of hardware, software and well trained ISS and IT officers. To start the process, the Division should have the manpower structure as given below:

As proper coordination with line ministries and State Governments and other national and international agencies are involved, the Data Management Division should be managed by an officer at Spl. DG level with other well trained supporting Statistical Personnel namely 4 ADG, 6 DDG, 12 JAG, 24 STS/JTS level ISS Officers with suitable supporting Programmers, Database Administrators, System Analysts and other supporting staff. Their role is to provide all types of data say micro, macro, analytical data sets, time series data, and look after the proper keep up of data maintenance and updation at national level. This Division will have the additional responsibilities of data integration, matching and conversion of raw data into system independent portable data which could be used directly for further analyses. The starting point could be NSO data sets, followed by State data sets and then private sectors covering all segments of the economy. This Division could be expanded over a period of time while expanding its functioning.

Further, data should flow from State Agencies as well which could be used at National level for aggregation. Similar set-up is required at state level under the overall control of CSO. In each state the staff size required is 1 DDG, 2 JAG, 4 STS/JTS level ISS officers. Their role is to provide all types of data say micro, macro, analytical data sets, time series data and look after the proper keep up of data maintenance and updation at state level.

Chapter - 2

Conceptual framework on Data management

Data management

2.1 Introduction

Indian Statistical System generates enormous amount of data through multitude of agencies. The data are to be harmonized, integrated and made available through single window system to all types of users, say planners, policy makers, researcher and citizen. Ministry of Statistics, other central ministries and departments of the Union Government, State governments are engaged in the collection, compilation and dissemination of official Statistics. Many of the data series are a by-product of the general administration of the Ministries and States based on the records of the concerned offices, as also a product of the administration of particular Acts of the Government and Rules framed under them. For any developmental attempt to harmonize, integrate and make available the timely, credible and reliable data sets to the users, it is essential to know the conceptual frame work of "data management". This chapter mainly deals with fundamental concepts, standards and implementation procedures while developing "Data Management" for a huge country like India.

Data management is the development and execution of architectures, policies, practices and procedures in order to manage the information lifecycle needs of an organization in an effective manner.

Data Management is an overarching term that refers to all aspects of creating, housing, delivering, maintaining and retiring data that today adds the new contexts of compliance and the goal of managing data as an organisational asset. Data management typically addresses the creation of data architecture and is inclusive of the infrastructure, personnel, processes and other requirements for identifying, consolidating and optimizing data assets for efficiency and usefulness. Increasingly, data management falls under the rubric of data governance, a structured and role-oriented methodology for delivering dependable data assets for business decision support.

Data management comprises all the disciplines related to managing data as a valuable resource. The general definition is: "Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise/organization." This definition is fairly broad and encompasses a number of professions which may not have direct technical contact with lower-level aspects of data management, such as relational database management.

Alternatively, the definition provided in the Data Management Body of Knowledge (DMBOK) is: "Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets."

The concept of "Data Management" arose in the 1980s as technology moved from sequential processing (first cards, then tape) to random access processing since it was technically possible to store a single fact in a single place and access that using

random access disk. As applications moved more and more into real-time, interactive applications, it became obvious to most practitioners that data management is important. If the data was not well defined, the data would be misused in applications.

2.2 Disciplines in Data Management

Disciplines in Data Management include:

1. Data governance
    - Data asset
    - Data governance
    - Data steward
2. Data Architecture, Analysis and Design
    - Data analysis
    - Data architecture
    - Data modelling
3. Database Management
    - Data maintenance
    - Database administration
    - Database management system
4. Data Security Management
    - Data access
    - Data erasure
    - Data privacy
    - Data security
5. Data Quality Management
    - Data cleansing
    - Data integrity
    - Data quality
    - Data quality assurance
6. Reference and Master Data Management
    - Data integration
    - Master data management
    - Reference data
7. Data Warehousing and Business Intelligence Management
    - Business intelligence
    - Data mart
    - Data mining
    - Data movement (extract, transform and load)
    - Data warehousing
8. Document, Record and Content Management
    - Document management system
    - Records management
9. Meta Data Management
    - Meta-data management
    - Metadata
    - Metadata discovery
    - Metadata publishing

- Metadata registry

10. Contact Data Management
    - Business continuity planning
    - Marketing operations
    - User data integration
    - Identity management
    - Identity theft
    - Data theft
11. Data Movement(General)
12. Statistical Data and Metadata eXchange (SDMX)
13. Secured Data Centers(SDC)

2.3 Data Asset:

Data and information are the lifeblood of the 21st century economy. In the Information Age, data is recognized as a vital asset of an organisation. *"Organizations that do not understand the overwhelming importance of managing data and information as tangible assets in the new economy will not survive."* Data, and the information created from data, are now widely recognized as enterprise/organization assets.

No organization can be effective without high quality data. Today's organizations rely on their data assets to make more informed and more effective decisions. Market leaders are leveraging their data assets by creating competitive advantages through greater knowledge of their customers, innovative uses of information, and operational efficiencies. Businesses are using data to provide better products and services, cut costs, and control risks. Government agencies, educational institutions, and not-for-profit organizations also need high quality data to guide their operational, tactical, and strategic activities. As organizations need and increasingly depend on data, the business value of data assets can be more clearly established.

The amount of data available in the world is growing at an astounding rate. Researchers at the University of California at Berkeley estimate that the world produces between 1 and 2 trillion bytes of data annually. It often seems we are drowning in information.

Yet for many important decisions, we experience information gaps – the difference between what we know and what we need to know to make an effective decision. Information gaps represent Organization liabilities with potentially profound impacts on operational effectiveness and profitability.

Every organization needs to effectively manage its increasingly important data and information resources. Through a partnership of business leadership and technical expertise, the data management function can effectively provide and control data and information assets.

2.4 Data governance

Data governance is a quality control discipline for assessing, managing, using, improving, monitoring, maintaining, and protecting organizational information. It is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.

Data governance is a set of processes that ensures that important data assets are formally managed throughout the organisation. Data governance ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality. It is about putting people in charge of fixing and preventing issues with data so that the organisation can become more efficient. Data governance also describes an evolutionary process for an organization, altering the organization's way of thinking and setting up the processes to handle information so that it may be utilized by the entire government. It's about using technology when necessary in many forms to help aid the process. When organisation desire, or are required, to gain control of their data, they empower their people, set up processes and get help from technology to do it.

Data governance encompasses the people, processes, and information technology required to create a consistent and proper handling of an organization's data across the business enterprise. Goals may be defined at all levels of the organisation and doing so may aid in acceptance of processes by those who will use them. Some goals include:

- Increasing consistency and confidence in decision making
- Decreasing the risk of regulatory fines
- Improving data security
- Maximizing the income generation potential of data
- Designating accountability for information quality
- Enable better planning by supervisory staff
- Minimizing or eliminating re-work
- Optimize staff effectiveness
- Establish process performance baselines to enable improvement efforts
- Acknowledge and hold all gains

These goals are realized by the implementation of Data governance programs, or initiatives using Change Management techniques.

2.5 Data steward

In metadata, a **data steward** is a person that is responsible for maintaining a data element in a metadata registry. A data steward may share some responsibilities with a data custodian.

Data stewardship roles are common when organizations are attempting to exchange data precisely and consistently between computer systems and reuse data-related

resources. Master data management often makes references to the need for data stewardship for its implementation to succeed.

Data Steward Responsibilities

A data steward ensures that each assigned data element:

1. Has clear and unambiguous data element definition.
2. Does not conflict with other data elements in the metadata registry (removes duplicates, overlap etc.)
3. Has clear enumerated value definitions if it is of type Code.
4. Is still being used (remove unused data elements)
5. Is being used consistently in various computer systems
6. Has adequate documentation on appropriate usage and notes
7. Documents the origin and sources of authority on each metadata element

2.6 Data Architecture, Analysis and Design

Data analysis

Analysis of data is a process of inspecting, cleaning, transforming, and modeling **data** with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis(EDA), and confirmatory data analysis(CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical or structural models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis.

Data integration is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination. The term *data analysis* is sometimes used as a synonym for data modelling.

Type of data

Data can be of several types

- Quantitative data; data is a number

- Categorical data; data one of several categories
- Qualitative data; data is a pass/fail or the presence of a characteristic

Data cleaning

Data cleaning is an important procedure during which the data are inspected, and erroneous data are—if necessary, preferable, and possible—corrected. Data cleaning can be done during the stage of data entry. If this is done, it is important that no subjective decisions are made. The guiding principle is during subsequent manipulations of the data, information should always be cumulatively retrievable. In other words, it should always be possible to undo any data set alterations. Therefore, it is important not to throw information away at any stage in the data cleaning phase. All information should be saved (i.e., when altering variables, both the original values and the new values should be kept, either in a duplicate dataset or under a different variable name), and all alterations to the data set should carefully and clearly documented, for instance in a syntax or a log.

2.7 Data Architecture

Data Architecture in enterprise architecture is the design of data for use in defining the target state and the subsequent planning needed to achieve the target state. It is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.

A data architecture describes the data structures used by a business and/or its applications. There are descriptions of data in storage and data in motion; descriptions of data stores, data groups and data items; and mappings of those data artifacts to data qualities, applications, locations etc.

Essential to realizing the target state, Data Architecture describes how data is processed, stored, and utilized in a given system. It provides criteria for data processing operations that make it possible to design data flows and also control the flow of data in the system.

The Data Architect is responsible for defining the target state, alignment during development and then minor follow up to ensure enhancements are done in the spirit of the original blueprint.

During the definition of the target state, the Data Architecture breaks a subject down to the atomic level and then builds it back up to the desired form. The Data Architect breaks the subject down by going through 3 traditional architectural processes:

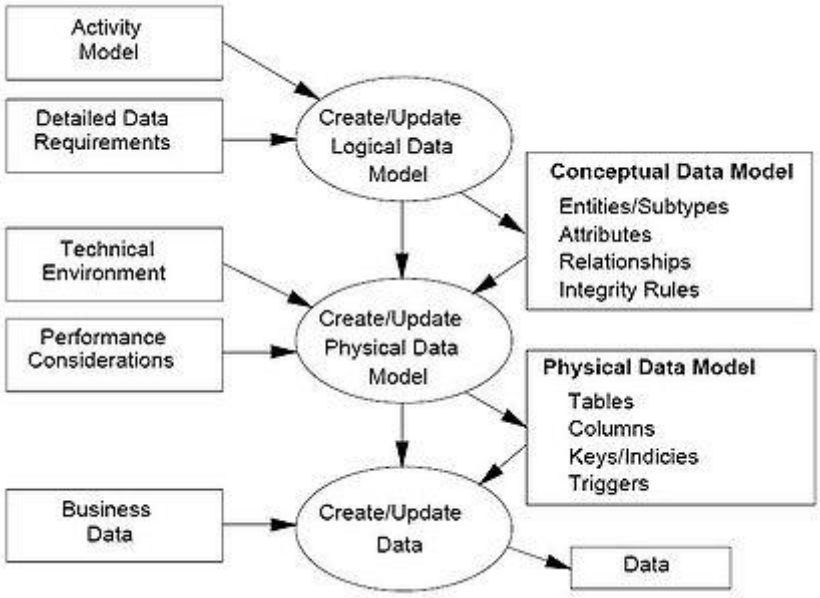
- Conceptual - represents all business entities.
- Logical - represents the logic of how entities are related.
- Physical - the realization of the data mechanisms for a specific type of functionality.

Data architecture includes a complete analysis of the relationships between an organization's functions, available technologies, and data types.

Data architecture should be defined in the **planning phase** of the design of a new data processing and storage system. The major types and sources of data necessary to support an organisation should be identified in a manner that is complete, consistent, and understandable. The primary requirement at this stage is to define all of the relevant **data entities**, not to specify computer hardware items. A data entity is any real or abstracted thing about which an organization or individual wishes to store data.

2.8 Data modelling

Data modelling in software engineering is the process of creating a data model by applying formal data model descriptions using data modelling techniques.



The data modelling process. The figure illustrates the way data models are developed and used today. A conceptual data model is developed based on the data requirements for the application that is being developed, perhaps in the context of an activity model. The data model will normally consist of entity types, attributes, relationships, integrity rules, and the definitions of those objects. This is then used as the start point for interface or database design.

Data modelling is a method used to define and analyze data requirements needed to support the business processes of an organization. The data requirements are recorded as a conceptual data model with associated data definitions. Actual implementation of the conceptual model is called a logical data model. To implement one conceptual data model may require multiple logical data models. Data modelling

defines not just data elements, but their structures and relationships between them Data modelling techniques and methodologies are used to model data in a standard, consistent, predictable manner in order to manage it as a resource. The use of data modelling standards is strongly recommended for all projects requiring a standard means of defining and analyzing data within an organization, e.g., using data modelling:

- to manage data as a resource;
- for the integration of information systems;
- for designing databases/data warehouses (aka data repositories)

Data modelling may be performed during various types of projects and in multiple phases of projects. Data models are progressive; there is no such thing as the final data model for a business or application. Instead a data model should be considered a living document that will change in response to a changing business. The data models should ideally be stored in a repository so that they can be retrieved, expanded, and edited over time.

Data modelling is also a technique for detailing business requirements for a database. It is sometimes called *database modelling* because a data model is eventually implemented in a database.

2.9 Database Management

2.10 Data maintenance

Data maintenance is the adding, deleting, changing and updating of binary and high level files, and the real world data associated with those files. Data can be maintained manually and/or through an automated program, but at origination and translation/delivery point must be translated into a binary representation for storage. Data is usually edited at a slightly higher level in a format relevant to the content of the data (such as text, images, or scientific or financial information). It is also the backing up, storage and general up keep of this all this data in the long term.

2.11 Database Administration

Database administration is extremely important in managing data. Every organization or enterprise needs database administrators that are responsible for the database environment. Database administrators are usually given the authority to do the following tasks that include recoverability, integrity, security, availability, performance and development & testing support.

Recoverability is usually defined as a way to store data as a back up and then test the back ups to make sure that they are valid. The task of integrity means that data that is pulled for certain records or files are in fact valid and have high data integrity. Data integrity is extremely important especially when creating reports or when data is used for analysis. If you have data that is deemed invalid, your results will be worthless.

Database security is an essential task for database administrators. For instance, database administrators are usually in charge of giving clearance and access to certain databases or trees in an organization. Another important task is availability. Availability is defined as making sure a database is up and running. The more up time, usually the higher level of productivity. Performance is related to availability, it is considered getting the most out of the hardware, applications and data as possible. Performance is usually in relation to an organizations budget, physical equipment and resources.

Finally, a database administrator is usually involved in database development and testing support. Database administrators are always trying to push te envelope, trying to get more use out of the data and add better performing and more powerful applications, hardware and resources to the database structure. A database that is administered correctly is not only a sign of competent database administrator, but it also means that all end users have a huge resource in the data that is available. This makes it easy to create reports, conduct analysis and make high quality decisions based on data that is collected and used within the organization.

2.12 Database management system

A database management system is the system in which related data is stored in an efficient and compact manner. "Efficient" means that the data which is stored in the DBMS can be accessed quickly and "compact" means that the data takes up very little space in the computer's memory. The phrase "related data" is means that the data stored pertains to a particular topic.

Specialized databases have existed for scientific, imaging, document storage and like uses. Functionality drawn from such applications has begun appearing in mainstream DBMS's as well. However, the main focus, at least when aimed at the commercial data processing market, is still on descriptive attributes on repetitive record structures.

Thus, the DBMSs of today roll together frequently needed services or features of attribute management. By externalizing such functionality to the DBMS, applications effectively share code with each other and are relieved of much internal complexity.

2.13 Data Security Management

2.14 Data access

Data access typically refers to software and activities related to storing, retrieving, or acting on data housed in a database or other repository. There are two types of data access, sequential access and random access.

Historically, different methods and languages were required for every repository, including each different database, file system, etc., and many of these repositories stored their content in different and incompatible formats.

In more recent days, standardized languages, methods, and formats, have been created to serve as interfaces between the often proprietary, and always

idiosyncratic, specific languages and methods. Such standards include SQL, ODBC, JDBC, ADO.NET, XML, XQuery, XPath, and Web Services.

Some of these standards enable translation of data from unstructured (such as HTML or free-text files) to structured (such as XML or SQL)

2.15 Data erasure

Data erasure (also called data clearing or data wiping) is a software-based method of overwriting data that completely destroys all electronic data residing on a hard disk drive or other digital media. Permanent data erasure goes beyond basic file deletion commands, which only remove direct pointers to data disk sectors and make data recovery possible with common software tools. Unlike degaussing and physical destruction, which render the storage media unusable, data erasure removes all information while leaving the disk operable, preserving IT assets and the environment.

Software-based overwriting uses a software application to write patterns of random meaningless data onto all of a hard drive's sectors. There are key differentiators between data erasure and other overwriting methods, which can leave data intact and raise the risk of data breach or spill, identity theft and failure to achieve regulatory compliance. Many data eradication programs also provide multiple overwrites so that they support recognized government and industry standards. Good software should provide verification of data removal, which is necessary for meeting certain standards.

To protect data on lost or stolen media, some data erasure applications remotely destroy data if the password is incorrectly entered. Data erasure tools can also target specific data on a disk for routine erasure, providing a hacking protection method that is less time-consuming than encryption.

2.16 Data privacy

Information privacy

Information privacy, or **data privacy** is the relationship between collection and dissemination of data, technology, the public expectation of privacy, and the legal and political issues surrounding them.

Privacy concerns exist wherever personally identifiable information is collected and stored - in digital form or otherwise. Improper or non-existent disclosure control can be the root cause for privacy issues. Data privacy issues can arise in response to information from a wide range of sources, such as:

- Healthcare records
- Criminal justice investigations and proceedings
- Financial institutions and transactions
- Biological traits, such as genetic material
- Residence and geographic records

The challenge in data privacy is to share data while protecting personally identifiable information. The fields of data security and information security design and utilize software, hardware and human resources to address this issue.

2.17 Data security

Data security is the means of ensuring that data is kept safe from corruption and that access to it is suitably controlled. Thus data security helps to ensure privacy. It also helps in protecting personal data.

Data Security Technologies

Disk Encryption

Disk encryption refers to encryption technology that encrypts data on a hard disk drive. Disk encryption typically takes form in either software (say, disk encryption software] or hardware (say, disk encryption hardware). Disk encryption is often referred to as on-the-fly encryption ("OTFE") or transparent encryption.

2.18 Data Quality Management

Data cleansing

Data cleansing and **data scrubbing** are two very different processes and must not be considered to be the same. '**Data Cleansing** is particularly the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant etc. parts of the data and then replacing, modifying or deleting this *dirty data*.

Other practical functions involved in Data Cleansing include:

- Formatting the data so that attributes such as Product, Brand Name, Manufacturer's Name, Part Number etc.
- Expansion of all abbreviations which can be identified other than the industry standard abbreviations such as PTFE, ASTM, ASME etc.
- Unit of Measure (UOM) standardization by either use a customer specific style guide, or Unilog's style guide.

After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by different data dictionary definitions of similar entities in different stores, may have been caused by user entry errors, or may have been corrupted in transmission or storage.

Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data.

Data quality

High quality data needs to pass a set of quality criteria. Those include:

- **Accuracy**: An aggregated value over the criteria of integrity, consistency and density
- **Integrity**: An aggregated value over the criteria of completeness and validity
- **Completeness**: Achieved by correcting data containing anomalies
- **Validity**: Approximated by the amount of data satisfying integrity constraints
- **Consistency**: Concerns contradictions and syntactical anomalies
- **Uniformity**: Directly related to irregularities
- **Density**: The quotient of missing values in the data and the number of total values ought to be known
- **Uniqueness**: Related to the number of duplicates in the data

2.19 Data integrity

Data integrity is data that has a complete or whole structure. All characteristics of the data including business rules, rules for how pieces of data relate, dates, definitions and lineage must be correct for data to be complete.

Per the discipline of data architecture, when functions are performed on the data the functions must ensure integrity. Examples of functions are transforming the data, storing the history, storing the definitions (Metadata) and storing the lineage of the data as it moves from one place to another. The most important aspect of data integrity per the data architecture discipline is to expose the data, the functions and the data's characteristics.

Data that has integrity is identically maintained during any operation (such as transfer, storage or retrieval). Put simply in business terms, data integrity is the assurance that data is consistent, certified and can be reconciled.

In terms of a database data integrity refers to the process of ensuring that a database remains an accurate reflection of the universe of discourse it is modelling or representing. In other words there is a close correspondence between the facts stored in the database and the real world it models.

2.20 Data quality

Data are of high quality "if they are fit for their intended uses in operations, decision making and planning". Alternatively, the data are deemed of high quality if they correctly represent the real-world construct to which they refer. Furthermore, apart from these definitions, as data volume increases, the question of *internal consistency* within data becomes paramount, regardless of fitness for use for any external purpose, e.g. a person's age and birth date may conflict within different parts of a

database. The first views can often be in disagreement, even about the same set of data used for the same purpose. This article discusses the concept as it related to business data processing, although of course other data have various quality issues as well.

2.21 Data quality assurance

Data quality assurance is the process of profiling the data to discover inconsistencies, and other anomalies in the data and performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve the data quality .

These activities can be undertaken as part of Data warehousing or as part of the Database administration of an existing piece of applications software.

2.22 Reference and Master Data Management

Data integration

Data integration involves combining data residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situations both commercial (when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example). Data integration appears with increasing frequency as the volume and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved. In management circles, people frequently refer to data integration as "Enterprise Information Integration" (EII).

2.23 Master data management

Master data management (MDM) comprises a set of processes and tools that consistently defines and manages the non-transactional data entities of an organization (which may include reference data). MDM has the objective of providing processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing such data throughout an organization to ensure

consistency and control in the ongoing maintenance and application use of this information.

The term recalls the concept of a *master file* from an earlier computing era. MDM is similar to, and some would say the same as, virtual or federated database management.

At a basic level, MDM seeks to ensure that an organization does not use multiple (potentially inconsistent) versions of the same master data in different parts of its operations, which can occur in large organizations. A common example of poor MDM is the scenario of a bank at which a customer has taken out a mortgage and the bank begins to send mortgage solicitations to that customer, ignoring the fact that the person already has a mortgage account relationship with the bank. This happens because the customer information used by the marketing section within the bank lacks integration with the customer information used by the customer services section of the bank. Thus the two groups remain unaware that an existing customer is also considered a sales lead.

Other problems include (for example) issues with the quality of data, consistent classification and identification of data, and data-reconciliation issues.

2.24 Reference data

Reference data are data describing a physical or virtual object and its properties. Reference data are usually described with nouns.

Reference data is used in data management to define characteristics of an identifier that are used within other data centric processes. For example - reference data within finance might be a product master or a security master.

Typical reference data are:

- Physical: products, material, assets, customers, locations
- Virtual: cost centers, planned buildings

Reference data can change over time via transactions (described in transaction data). E.g. a logistical transaction can change the location of an object, a financial transaction like adding tax can change the price of an object and a series of work transaction can change a virtual object like a planned building into a physical object.

Master reference data

A special type of reference data is **master reference data** - these are reference data shared over a number of systems. Some master reference data are universal (like the list of Countries) and can be covered by a global standard (in this case ISO).

Master data and Master reference data

Master reference data are sometimes called "Master Data". This usage of the term Master data should be avoided, since Master data is also the term used for **original data**, like an original recording.

Master data is collected on a Data Management Platform. The platform enables the acquisition, management and distribution of vendor and internally sourced Reference Data.

2.25 Data Warehousing and Business Intelligence Management

2.26 Business intelligence

Business intelligence (**BI**) refers to computer-based techniques used in spotting, digging-out, and analyzing data, such as sales revenue by products and/or departments, or by associated costs and incomes.

BI technologies provide historical, current, and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, business performance management, benchmarking, text mining, and predictive analytics.

Business intelligence aims to support better business decision-making. Thus a BI system can be called a decision support system (DSS).

Business intelligence and data warehousing

Often BI applications use data gathered from a data warehouse or a data mart. However, not all data warehouses are used for business intelligence, nor do all business intelligence applications require a data warehouse.

In order to distinguish between concepts of business intelligence and data warehouses, experts often defines business intelligence in one of two ways:

Typically, experts uses the following broad definition: "Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making." When using this definition, business intelligence also includes technologies such as data integration, data quality, data warehousing, master data management, text and content analytics, and many others that the market sometimes lumps into the Information Management segment. Therefore, experts refer to *data preparation* and *data usage* as two separate, but closely linked segments of the business intelligence architectural stack.

Experts define the latter, narrower business intelligence market as "referring to just the top layers of the BI architectural stack such as reporting, analytics and dashboards."

2.27 Data mart

A **data mart** (DM) is the access layer of the data warehouse (DW) environment that is used to get data out to the users. The DM is a subset of the DW, usually oriented to a specific business line or team.

In practice, the data mart and data warehouse each tend to imply the presence of the other in some form. However, most writers using the term seem to agree that the design of a data mart tends to start from an analysis of user needs and that a data warehouse tends to start from an analysis of what data already exists and how it can be collected in such a way that the data can later be used. A data warehouse is a central aggregation of data (which can be distributed physically); a data mart is a data repository that may or may not derive from a data warehouse and that emphasizes ease of access and usability for a particular designed purpose. In general, a data warehouse tends to be a strategic but somewhat unfinished concept; a data mart tends to be tactical and aimed at meeting an immediate need.

There can be multiple data marts inside a single corporation; each one relevant to one or more business units for which it was designed. DMs may or may not be dependent or related to other data marts in a single corporation. If the data marts are designed using conformed facts and dimensions, then they will be related. In some deployments, each department or business unit is considered the *owner* of its data mart including all the *hardware*, *software* and *data*. This enables each department to use, manipulate and develop their data any way they see fit; without altering information inside other data marts or the data warehouse. In other deployments where conformed dimensions are used, this business unit ownership will not hold true for shared dimensions like customer, product, etc.

Design schemas

- Star schema or dimensional model is a fairly popular design choice, as it enables a relational database to emulate the analytical functionality of a multidimensional database.
- Snowflake schema
- Datamart Architecture Pattern

2.28 Data Mining

Another important topic regarding data management is data mining. Data mining is a process in which large amounts of data are sifted through to show trends, relationships, and patterns. Data mining is a crucial component to data management because it exposes interesting information about the data being collected. It is important to note that data is primarily collected so it can be used to find these patterns, relationships and trends that can help a business grow or create profit.

While there are many topics within data management, they all work together from the beginning where data is collected to the end of the process where it is sifted through; analyzed and formatted where specialists can then make quality decisions based upon it

2.29 Data movement (extract, transform and load)

Extract, transform and load (**ETL**) is a process in database usage and especially in data warehousing that involves:

- Extracting data from outside sources
- Transforming it to fit operational needs (which can include quality levels)
- Loading it into the end target (database or data warehouse)

ETL Architecture Pattern

Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization/format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even fetching from outside sources such as through web spidering or screen-scraping. The streaming of the extracted data source and load on-the-fly to the destination database is another way of performing ETL when no intermediate data storage is required. In general, the goal of the extraction phase is to convert the data into a single format which is appropriate for transformation processing.

Transform

The transform stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the target database:

Load

The load phase loads the data into the end target, usually the data warehouse (DW). Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information, frequently updating extract data is done on daily, weekly or monthly basis. Other DW (or even other parts of the same DW) may add new data in a historicized form, for example, hourly. To understand this, consider a DW that is required to maintain sales records of the last year. Then, the DW will overwrite any data that is older than a year with newer data. However, the entry of data for any one year window will be made in a historicized manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the DW.

2.30 Data Warehousing

Data warehousing is storing data effectively so that it can be accessed and used efficiently. Different organizations collect different types of data, but many organizations use their data the same way, in order to create reports and analyze their data to make quality business decisions. Data warehousing is usually an organizational wide repository of data, however for very large corporations in can encompass just one office or one department.

Data Warehouse Intelligence is a term to describe a system used in an organization to collect data, most of which are transactional data, such as purchase records and etc., from one or more data sources, such as the database of a transactional system, into a central data location, the Data Warehouse, and later report those data, generally in an aggregated way, to business users in the organization. This system generally consists of an ETL tool, a Database, a Reporting tool and other facilitating tools, such as a Data Modelling tool.

A **data warehouse** (DW) is a database used for reporting. The data is offloaded from the operational systems for reporting. The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

A data warehouse maintains its functions in three layers: staging, integration, and access. *Staging* is used to store raw data for use by developers (analysis and support). The *integration* layer is used to integrate data and to have a level of abstraction from users. The *access* layer is for getting data out for users.

2.31 Document, Record and Content Management

2.32 Document management system

A **document management system** (DMS) is a computer system (or set of computer programs) used to track and store electronic documents and/or images of paper documents. It is usually also capable of keeping track of the different versions created by different users (history tracking). The term has some overlap with the concepts of content management systems. It is often viewed as a component of enterprise content management (ECM) systems and related to digital asset management, document imaging, workflow systems and records management systems.

Document management systems commonly provide storage, versioning, metadata, security, as well as indexing and retrieval capabilities.

2.33 Records management

Records management, or **RM**, is the practice of maintaining the records of an organization from the time they are created up to their eventual disposal. This may

include classifying, storing, securing, and destruction (or in some cases, archival preservation) of records.

A record can be either a tangible object or digital information: for example, birth certificates, medical x-rays, office documents, databases, application data, and e-mail. Records management is primarily concerned with the evidence of an organization's activities, and is usually applied according to the value of the records rather than their physical format.

Electronic records management systems

An Electronic Document and Records Management System (EDRM) is a computer program (or set of programs) used to track and store records. The term is distinguished from imaging and document management systems that specialize in paper capture and document management respectively. ERM systems commonly provide specialized security and auditing functionality tailored to the needs of records managers.

2.34 Meta Data Management

2.35 Meta-data management

Meta-data management (also known as metadata management, without the hyphen) involves storing information about other information. With different types of media being used, references to the location of the data can allow management of diverse repositories.

URLs, images, video etc. may be referenced from a triples table of object, attribute and value.

With specific knowledge domains, the boundaries of the metadata for each must be managed, since a general ontology is not useful to experts in one field whose language is knowledge-domain specific.

2.36 Metadata

The term **Metadata** is an ambiguous term which is used for two fundamentally different concepts (Types). Although a trite expression "data about data" is often used, it does not apply to both in the same way. Structural metadata, the design and specification of data structures, cannot be about data, because at design time the application contains no data. In this case the correct description would be "data about the containers of data". Descriptive metadata on the other hand, is about individual instances of application data, the data content. In this case, a useful description (resulting in a disambiguating neologism) would be "data about data contents" or "content about content" thus Metacontent. Descriptive, Guide and the NISO concept of Administrative metadata are all subtypes of metacontent.

Metadata is data. As such, metadata can be stored and managed in a database, often called a registry or repository. However, it is impossible to identify metadata

just by looking at it because a user would not know when data is metadata or just data.

2.37 *Metadata discovery*

In metadata, **metadata discovery** is the process of using automated tools to discover the semantics of a data element in data sets. This process usually ends with a set of mappings between the data source elements and a centralized metadata registry.

Metadata discovery is also known as metadata scanning.

2.38 Metadata publishing

Metadata publishing is the process of making metadata data elements available to external users, both people and machines using a formal review process and a commitment to change control processes.

Metadata publishing is the foundation upon which advanced distributed computing functions are being built. But like building foundations, care must be taken in metadata publishing systems to ensure the structural integrity of the systems built on top of them.

Published metadata has the following characteristics:

1. Metadata structures available to the general public on a public web site or by a download
2. There is a documented review and approval process for adding or updating data elements to the system
3. New releases are made available without disturbing prior versions
4. A publishing organization that makes a commitment to change control process

2.39 Metadata registry

A **metadata registry** is a central location in an organization where metadata definitions are stored and maintained in a controlled method.

Use of Metadata Registries

Metadata registries are used whenever data must be used consistently within an organization or group of organizations. Examples of these situations include:

- Organizations that transmit data using structures such as XML, Web Services or EDI
- Organizations that need consistent definitions of data across time, between databases, between organizations or between processes, for example when an organization builds a data warehouse

- Organizations that are attempting to break down "silos" of information captured within applications or proprietary file formats

Central to the charter of any metadata management programme is the process of creating trusting relationships with stakeholders and that definitions and structures have been reviewed and approved by appropriate parties.

2.40 Contact Data Management

2.41 Business continuity planning

Business continuity planning life cycle

Business continuity planning (**BCP**) is "planning which identifies the organization's exposure to internal and external threats and synthesizes hard and soft assets to provide effective prevention and recovery for the organization, whilst maintaining competitive advantage and value system integrity". It is also called **Business continuity & Resiliency planning** (**BCRP**). The logistical plan used in BCP is called a ***business continuity plan***. The intended effect of BCP is to ensure business continuity, which is an ongoing state or methodology governing how business is conducted.

2.42 Marketing operations

Typically, Marketing Operations is the function responsible for marketing performance measurement, strategic planning and budgeting, process development, professional development, and marketing systems and data. This work either connects closely to, or includes, demand generation. It also involves the alignment of Marketing with Sales, Business Units, and Finance. Marketing Operational professionals are not classical marketers. Instead of coming from PR or branding backgrounds, they typically come from Finance, IT, Sales Operations and other analytical or process-oriented roles.

2.43 User data integration (U**DI**)

In data processing, **User data integration** (U**DI**) combines the technology, processes and services needed to set up and maintain an accurate, timely, complete and comprehensive representation of a user across multiple channels, business-lines, and enterprises — typically from multiple sources of associated data in multiple application systems and databases. It applies data-integration techniques in this specific area.

UDI commonly supports both user relationship management and master data management, and enables access from these organisation applications to information confidently describing everything known about a user, donor, or prospect, including all attributes and cross references, along with the critical definition and identification necessary to uniquely differentiate one user from another and their individual needs.

2.44 Identity management

Identity management (or **ID management**, or simply **IdM**) is a broad administrative area that deals with identifying individuals in a system (such as a country, a network, or an organization) and controlling access to the resources in that system by placing restrictions on the established identities of the individuals.

Identity management is multidisciplinary and covers many dimensions, such as:

- Technical – Employs identity management systems (identification, implementation, administration and termination of identities with access to information systems, buildings and data within an organization).
- Legal – Deals with legislation for data protection.
- Police – Deals with identity theft.
- Social and humanity – Deals with issues such as privacy.
- Security – Manages elements such as access control.
- Organizations – Hierarchies and divisions of access.

Identity management (IdM) is a term related to how humans are identified and authorized across computer networks. It covers issues such as how users are given an identity, the protection of that identity, and the technologies supporting that protection (e.g., network protocols, digital certificates, passwords, etc.).

Digital identity: Personal identifying information (PII) selectively exposed over a network. See OECD and NIST guidelines on protecting PII and the risk of identity theft.

2.45 Identity theft

Identity theft is a form of fraud or cheating of another person's identity in which someone pretends to be someone else by assuming that person's identity, typically in order to access resources or obtain credit and other benefits in that person's name. The victim of identity theft (here meaning the person whose identity has been assumed by the identity thief) can suffer adverse consequences if he or she is held

accountable for the perpetrator's actions. Organizations and individuals who are duped or defrauded by the identity thief can also suffer adverse consequences and losses, and to that extent are also victims.

2.46 Data theft

Data theft is a growing problem primarily perpetrated by office workers with access to technology such as desktop computers and hand-held devices capable of storing digital information such as USB flash drives, iPods and even digital cameras. Since employees often spend a considerable amount of time developing contacts and confidential and copyrighted information for the company they work for they often feel they have some right to the information and are inclined to copy and/or delete part of it when they leave the company, or misuse it while they are still in employment.

While most organizations have implemented firewalls and intrusion-detection systems very few take into account the threat from the average employee that copies proprietary data for personal gain or use by another company. A common scenario is where a sales person makes a copy of the contact database for use in their next job. Typically this is a clear violation of their terms of employment.

The damage caused by data theft can be considerable with today's ability to transmit very large files via e-mail, web pages, USB devices, DVD storage and other hand-held devices. Removable media devices are getting smaller with increased hard drive capacity, and activities such as pod-slurping are becoming more and more common. It is now possible to store more than 160 GB of data on a device that will fit in an employee's pocket, data that could contribute to the downfall of a business.

2.47 Data Movement

Data movement is the ability to move data from one place to another. For instance, data needs to be moved from where it is collected to a database and then to an end user, but this process takes quite a bit of logistic insight. Not only do all hardware, applications and data collected need to be compatible with one another, they must also be able to be classified, stored and accessed with ease within an organization. Moving data can be very expensive and can require lots of resources to make sure that data is moved efficiently, that data is secure in transit and that once it reaches the end user it can be used effectively either to be printed out as a report, saved on a computer or sent as an email attachment.

2.48 Statistical Data and Metadata eXchange (SDMX)

Core tasks of National Statistical Organisation are to collect, process and organise statistical data, and subsequently put them at the disposal of various communities of users, often termed as dissemination of the statistics. Obviously, some of the main obligations of NSOs are to make the necessary strategy decisions on what should be measured and how, and to manage and document the statistical system.

A widespread problem is lack of harmonisation across different fields of statistics in a

country, even within the same national organisation. This is often related to the statistics production being organised in so-called stove-pipes, or independent production lines. This makes it difficult to use statistics for different subjects in a coherent way, thus impairing the quality of statistics as seen from the user perspective. It also reduces efficiency in the production process.

To overcome these problems, there has been a strong tendency in NSOs towards standardisation and integration, breaking down stove-pipes. This leads to the creation of corporate statistical data warehouses, bringing together statistics on different subjects under one system. In this endeavour, the creation of statistical metadata plays an important part. The changes required towards such integrated systems are not only technical, but also organisational.

The Statistical Data and Metadata eXchange (SDMX) initiative was launched in 2001 by seven organisations working on statistics at the international level: the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistical Division (UNSD) and the World Bank. These seven organisations act as the sponsors of SDMX. The stated aim of SDMX was to develop and use more efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries.

To achieve this goal, SDMX provides standard formats for data and metadata, together with content guidelines and an IT architecture for exchange of data and metadata. Organisations are free to make use of whichever elements of SDMX are most appropriate in a given case. With the Internet and the world-wide web, the electronic exchange and sharing of data has become easier and more common, but the exchange has often taken place in an *ad hoc* manner using all kinds of formats and non-standard concepts. This creates the need for common standards and guidelines to enable more efficient processes for exchange and sharing of statistical data and metadata. As statistical data exchange takes place continuously, the gains to be realised from adopting common approaches are considerable both for data providers and data users.

SDMX aims to ensure that metadata always come along with the data, making the information immediately understandable and useful. For this reason, the SDMX standards and guidelines deal with both data and metadata. Common standards and guidelines followed by all players not only help to give easy access to statistical data, wherever these data may be and without demanding prior agreement between two partners, but they also facilitate access to metadata that make the data more comparable, more meaningful and generally more usable.

The SDMX standards are designed for exchange or sharing of statistical information between two or more partners. Evidently, the SDMX standards have been developed by the sponsoring organisations in order to accommodate their constituencies, which include national statistical offices, central banks, ministries and other bodies. Within and across these constituencies, the standards are intended for reporting or sharing statistical data and metadata in the most efficient way.

SDMX standards can also be used within a national system for transmitting or sharing statistical data and metadata. This is particularly interesting in countries with a federal structure or a fairly decentralised statistical system. In such cases, a close link can be established between the national system for data sharing and the international ones, allowing for additional efficiency gains for the involved organisations. The use of SDMX for data exchange can easily evolve towards open SDMX-based dissemination; such dissemination may respond well to user demands for well structured data and metadata in reusable formats, and should be considered as an option for national authorities as well as international organisations. It is also an interesting option for private data providers, such as re-sellers of statistical databases.

SDMX can also be used for data and metadata management *within* statistical organisations, since its information model is applicable for much of the information stored and processed within statistical organisations, and such organisations can make use of the SDMX IT tools to reduce the costs of developing their data management systems.

The data (and related metadata) for a particular statistical domain are structured according to a "*Data Structure Definition"* (DSD, formerly known as a "key family"). The DSD describes the structure of a particular statistical data flow through a list of dimensions (for example: country, variable/topic, year), a list of "attributes" (for example, unit of measure) and their associated code lists. Attributes are metadata about an individual value, a time series or a group of time series.

SDMX also defines a model for additional explanatory metadata, which are often referred to in SDMX as *reference metadata.* Reference metadata are generally in a textual format, using concepts describing the content, methodology and quality of the data. The reference metadata for a particular statistical data flow, statistical domain or – if used homogeneously throughout a statistical institution - a particular statistical institution are structured according to a *"Metadata Structure Definition"* (MSD).

The *data exchange process* is represented in SDMX via the definition of "data flows", "data providers", and, in particular "provision agreements". The latter describes the way in which data and metadata are provided by a data provider. Thus, a data provider can express the fact that it provides a particular data flow covering a specific set of countries and topics, with a particular publication schedule. India should be part of this programme.

2.49 Secured Data Center (SDC)

The Secure Data Center (SDC) is a new service, intended to promote excellence in research by enabling safe and secure remote access by bona fide researchers to data hitherto deemed too sensitive, detailed, confidential or potentially disclosive to be made available under standard licensing and dissemination arrangements.

The Secure Data Center (SDC) gives researchers on-campus access to sensitive and confidential statistics. The space provides researchers with an area to work on secure data, in addition to providing a secure server room for data that needs to be physically inaccessible. The goal of the Center is to meet the increasing demand for access to the growing availability of microdata, but within a highly secure setting.

The SDC is unique in its approach to safe remote access. Its security philosophy is based upon training and trust, underpinned by cutting edge technology, appropriate licensing and legal frameworks, and strict security policies and penalties.

Another important potential role for the SDS is to provide a secure setting where appropriately authorised researchers can work with survey data matched with potentially disclosive administrative data (such as educational records), particularly for those researchers whose home institutions cannot provide the rigorous IT security environment required to receive such linked data.

Registration and access

The user will be required to complete the Accredited Researcher application. Once approval has been received, the user and a senior member of staff at their institution sign the SDC Service Agreement. The user is also required to have attended SDC training within the past stipulated period. Subject to these requirements, the user is given a service-specific user ID and password, which they can then use to access the service from their home institution.

Security philosophy

During its development phase for the pilot service, SDC staff spoke with a number of secure data enclaves around the world, to find out how and why security breaches occur. It became clear that the weakest link is not the technology, nor the data handling and procedural issues. Rather the potential for a security breach lies, as always, in human beings.

In talking to our sister services, it became clear that the tabloid image of the intentional malicious stealing and disclosure of personal data by some kind of master techno-criminal is simply not a real-life issue in the experience of secure data enclaves.

In real-life breaches, there have essentially been two kinds: one stems from ignorance of proper statistical disclosure control and data handling, and the other stems from a user's desire to escape the limitations of restricted onsite access for convenience's sake.

The SDC philosophy is to attack these two weaknesses directly: one with requiring SDC users to undergo explicit training on data security, data handling, statistical disclosure control and the use, where appropriate, of disclosure assessment tools before they can gain access to the system; and the other by creating a positive comfortable 'home away from home' analytical environment accessible at their convenience from the user's institution, if not their desktop, so that they will have no motivation to smuggle data out for use at home.

We maintain that providing remote secure access, far from being riskier than onsite data enclaves, actually *increases* service security, by removing the most common motivation for breaches. India should set up SDC as early as possible.

Chapter -3

Recommendations:

Recommendations

3.1 Introduction: In India, enormous amount of data is generated by the Central and State governments. Though Ministry of Statistics and Programme Implementation is declared as nodal agency for statistical activities, without sufficient manpower and authority, it is increasingly becoming difficult for the Ministry to exercise its mandate fully. Every line Ministry is functioning on its own-will in statistical matter resulting in lack of uniformity and standardization (ad-hoc mechanism) in collection and dissemination of data. At times MOSPI is not in a position to meet the demands of the line Ministry when the later aspires for conduct of surveys on subject most relevant and dear to them. Even MOSPI is not able to serve its own requirement for want of man power. Data collection for New Series of CPI (is outsourced) and various regular surveys are being carried out through contract employees. Issues are highly complicated, data requirements are immensely challenging and public criticism are mounting day by day.

In the light of the above observations, the recommendations are grouped into two main sections. The first one deals with continuing with the existing set up and integrating the data sources so as to provide data to the users through single window system. The second one deals with long term activity with the application of powerful IT tools integrating the entire gamut of statistics into single data warehouse server embedded with OLAP and OLAM Architectures to provide data through single window system adopting uniform format of data collection and dissemination system. The second one may be an ambitious goal but not difficult to achieve.

(3.2) Recommendations to be taken up immediately (short term)

1. Data availability discussed in the annexure-2 (Current national statistical system on key sectors) and issues discussed in the Chapter-1 may be the starting point. In every Ministry's web-site lot of information, data, reports, circulars, orders etc., are heavily loaded and some are up-to-date and some are not at all updated. All data sets are to be identified and converted into portable format(easy transfer of data set from one platform to another for further analyses and processing).

2. MOSPI being the nodal agency should enhance its resources in all angles like procurement and installation of latest Hardware & Software and staffing of well trained Technical manpower to meet the requirement.

3. MOSPI should transform all its data set to a portable data set format.

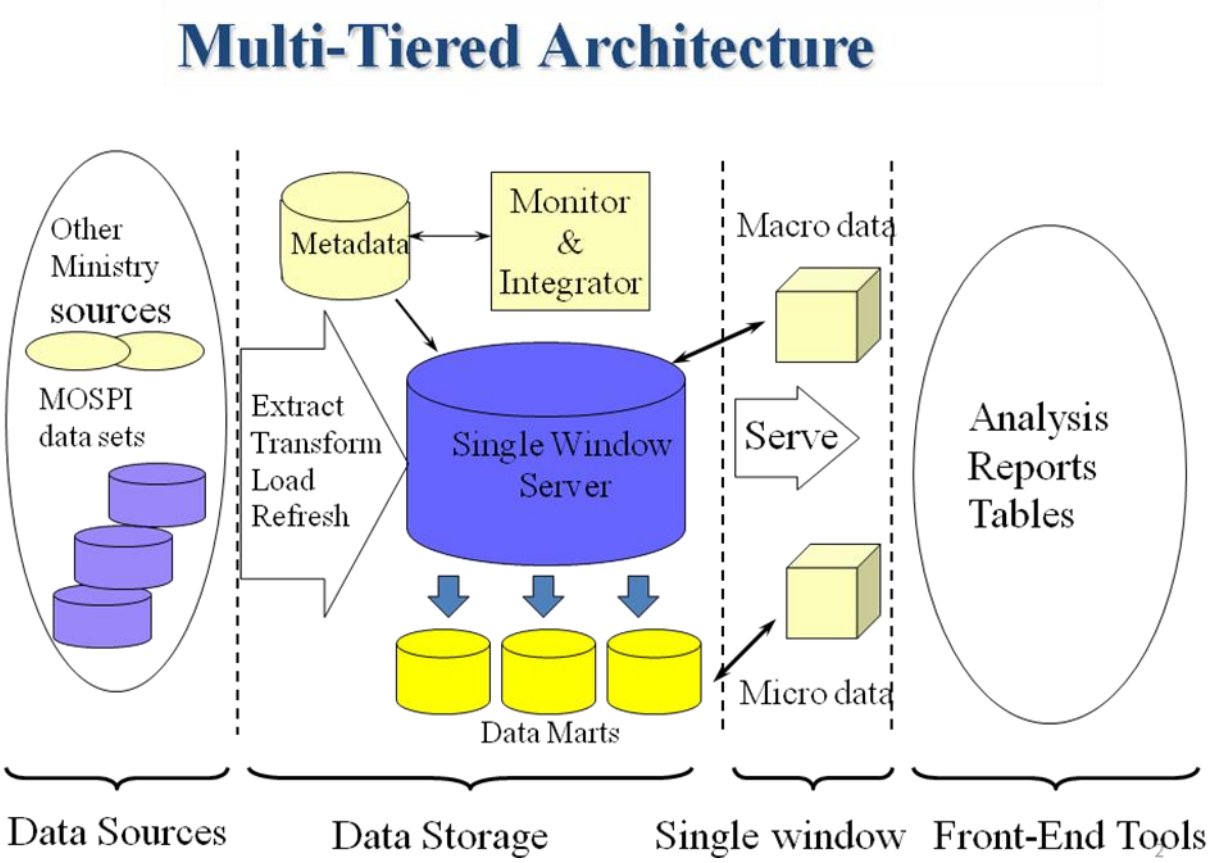
4. All data sets generated by MOSPI, line Ministries and State Governments should be loaded into a main server with all meta data details using ETL tools.

5. There should be a mechanism to update and load the data as and when new data set arrives and also a national policy on archiving the old data set(s).

6. Data set should be split into micro data and macro data set stream if need be two independent servers for each category. For example, (in the case of NSSO-Socio-Economic Surveys, Annual Surveys of Industries, etc. unit level data is made available after suppressing identification particulars) micro level data could be provided to users on request after suppressing the identification particulars thus ensuring that the confidentiality part is not compromised.

7. For simple analyses, Table generation and Report generation suitable and compatible software packages should be integrated into the system to the benefit of all types of users.

The complete process is presented in the form of simple and easily understandable multi-tiered Architecture scheme. Many conceptual portions have already been dealt in the chapter-2.



(3.3) Recommendations to be implemented over a period of time (Long term)

Though MOSPI had been a pioneer in introducing Computer and its application in its statistical activities in early sixties, it could not keep pace with the ever and fast changing technology due to various barriers and constraints. There is an urgent need to reverse this trend. ICT tools need to be extensively made use of right from data collection to data dissemination stage. Lot of hi-end propriety and free software are available. MOSPI should take advantage of these technological developments and integrate IT application and usage in the national statistical system.

Keeping these in view, the recommendations are being framed. This section is to recommend state-of-the-art IT application and integration of IT solutions in the national official statistical system. There should not be constraints and excuses on the premises of resources, manpower and changing technical and technological adoption. Though the recommendations are essential, implementations are time consuming due to various factors like attitudinal overhaul, approaches, commitments evaluation and migration from traditional set-up to technology driven set-up. This approach integrates hi-end data warehousing and data mining solutions embedded with online analytical processing (OLAP) and online analytical mining (OLAM) tools.

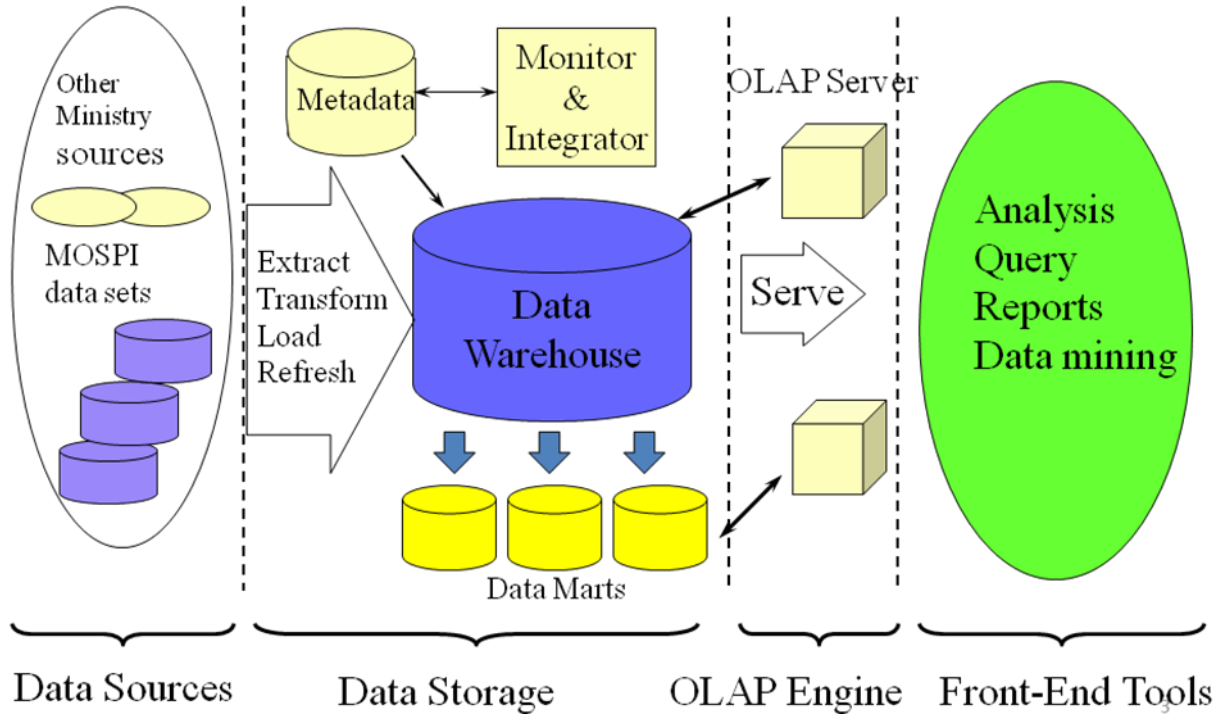
The following are recommended:

1. First and foremost requirement in case of decentralized data generation process is to define and adopt a uniform data format right from grass root level. It could be lowest level say village level or ultimate micro level like household or an individual to country level, data set should be uniform, single data format and easily portable. Various proprietary and free software are available which could be made use of for the purpose.

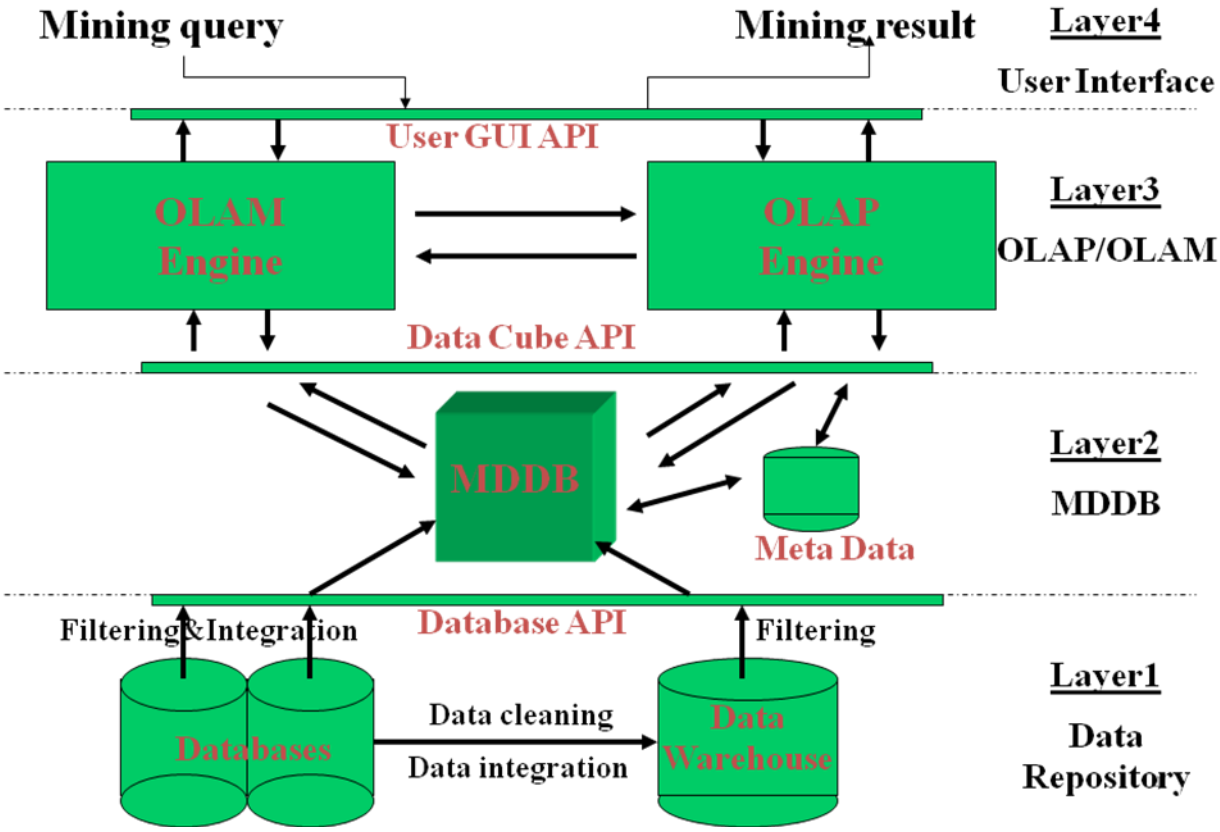
2. After having defined and standardized data set from different ministries, state governments and other data producing agencies, integrate (them and) transform and load them into centrally managed data warehouse server. Here application of ETL tools are necessary to transform the different data sets and load into compatible data warehousing server. Meta data and data-marts would also be taken into account.

3. OLAP and OLAM servers should be built on the nation wide data warehouse to enable the users to have analyses, query based filtering mechanisms, generate tables and reports and go into data mining solutions. The entire three steps approach is depicted in the following two figure, data warehouse architecture and OLAM Architecture.

Data warehouse Architecture



An OLAM Architecture



Entire concepts have been dealt already and hence they are not repeated here. The basic requirement is dedicated team of officers well versed with technology and statistics, cooperation of all line Ministries and of-course dedicated connectivity among MOSPI, line Ministries, State Governments and other data producers. User would be able to access all types of data say, micro, macro and derived data from single window.

1. Computer aided telephonic interview(CATI) and computer aided personal interview(CAPI) are very common method of data collection among the developed countries during the last 15 to 20 years. MOSPI should initiate data collection process in respect of all the surveys it conducts using computer assisted (computerized) data collection process. This helps in not only speedy data processing but minimize human errors at various stages of data collection and processing.

2. Evaluate and implement of end-to-end automation solution in the data process of NAS, NSSO-SE Surveys (collection to release of reports), Economic Censuses(collection to release of reports), IIP, ASI(collection to release of reports), and AS(collection to release of reports), by suitably deploying HW, SW and skilled manpower. Good example is CPI (New Series)

3. Set up Secured data centers to access confidential data by the authorized user. The data type could be sensitive like tax data, banking data, criminal records, etc. All required permission should be taken from the respective data producers to give access to licensed user in a secured environment. This would enable the licensed user to do all sorts of data analyses without taking the raw data.

4. Subscribe and take part in Statistical Data and Metadata eXchange (SDMX) programme. A widespread problem in data management in country like India is lack of harmonisation across different fields of statistics in a country, even within the same national organisation. This is often related to the statistics production being organised in so-called stove-pipes, or independent production lines. To overcome these problems, there has been a strong tendency in NSOs towards standardisation and integration, breaking down stove-pipes. This leads to the creation of corporate statistical data warehouses, bringing together statistics on different subjects under one system. The stated aim of SDMX is to develop and use more efficient processes for exchange and sharing of statistical data and metadata among national, international organisations and their member countries. The SDMX standards are designed for exchange or sharing of statistical information between two or more partners.

5. To implement the data management project of this mammoth dimension at national level, it may be appropriate to create a new division in CSO with dedicated composition of requisite manpower and other resources (H/W, S/W, Dedicated WAN (Optical Fiber Networks, etc.). As proper coordination with line ministries and State Governments and other national and international agencies are involved, the Data Management Division should be managed by an ISS officer at Spl. DG level with other well trained supporting Statistical Personnel namely 4 ADG, 6 DDG, 12 JAG, 24 STS/JTS level ISS Officers with suitable supporting Programmers, Database Administrators, System Analysts and other supporting staff. In each state the staff size required is 1 DDG, 2 JAG, 4 STS/JTS level ISS officers. This Division could be expanded over a period of time while expanding its functioning.

Annexures

(1) Minutes of the meetings of the Committee on Data Management

(2) Current national statistical system on key sectors

(3) National Data Dissemination Policy (1998)

Annexure-1a

(i) Minutes of the first meeting of the Committee

Minutes of the first meeting of the Committee on DATA MANAGEMENT

The first meeting of the Committee on DATA MANAGEMENT was held at 11-30 AM on 27th September, 2010 in the Committee Room (Room No. 223), II Floor, Sardar Patel Bhavan, New Delhi under the Chairmanship of Shri Suman Bery, Member, NSC. He welcomed the Members of the Committee and initiated the discussions. The list of participants is given in the annex.

After deliberating the importance of data collection, data management and user friendly data dissemination, Chairman referred the terms of reference of the Committee to all the members. Then he requested the DDG, NSC to inform the members on the objectives of the Committee.

DDG (NSC) has informed the Committee the various issues involved in data collection, data management and dissemination besides protocol to be followed in official statistical system. He expressed that data management scheme should provide easy access to the users and time frame should be fixed to the Data collection agencies for providing data for planners and users on completion of surveys and census. After the introduction of RTI Act 2005, as a matter of principal the data should be furnished without prolonged delay. Even Unit level data should be provided after suppressing identification particulars to maintain the confidentiality. Further, he informed that many Ministries and the organisations are having their web sites which are not user friendly. Besides these websites could be integrated and made available to the users through institutional mechanism. He also informed that the recommendations of the Committee would be used by NSC and Government for drawing national data management and dissemination policy.

Then, Chairman discussed the importance of timely availability of data, data integration, data customisation and inclusion of private sectors in data availability. Non-official Unit/Organisation like NCAER would participate in managing and providing processed data. He also informed that what RBI follows to develop and improve data management and data availability could be followed in official statistical system. Then, he invited the members to express their views on the subject matter.

Executive Director, RBI informed that scope and coverage of data collection mechanism should be enhanced. CSO and the NSSO have certain responsibilities in providing timely, reliable and quality data to the user community. The data dissemination system should be user friendly besides providing time series and historical data. Further he has informed that RBI has the responsibility of providing monetary statistics. RBI has deployed data warehousing and OLAP tools for analytical purposes. It is in the process of data integration and automation of transaction level data so that advanced data Mining tools can be deployed on the Data Warehouse platform. In financial sector including banking, data does not reside at the single place. RBI consolidates primary data and secondary data at national

level and produce the consolidated data through various publications (monthly, quarterly and annual). Statistics Handbook is updated on a near real time basis. RBI also supplies statistical products through CDs. Cross section data, integrating statistics division and economic division of RBI are produced. Further he has informed that the integrity of data above branch level is maintained.

Following the RBI views, the Chairman stressed the following three issues:

I **National best practices in providing data like advanced countries**: Focus should be on use of Information Technology in the best possible manner for data management on the lines of advanced countries like US, Canada, etc. to maintain and provide data including use of analytical techniques like data warehousing and data mining and dissemination should be at the multiple levels.

II **System of quality check**: Chairman enquired about the system check on quality of data attempted and deployment of mechanical process to check the quality. ED, RBI informed that the agencies have primary responsibilities in maintaining the quality of data. Further, he said that the meta-data details are not made available for other agencies by RBI.

III **Availability of meta data**. For integration of data, the basic requirement is availability of meta data. This also helps in maintaining cross links.

DDG, CSO (NAD) has informed that Ministry of Finance is the nodal agency for maintaining financial data. They have to ensure that all concerned are complying with international standards. GFS, BOP, latest Trade account and Capital accounts are yet to be compiled with international standards. These are in the process of compilation.

Representative of EPW has informed the committee that on financial statistics there is a serious data gap. No data on general government is available. RBI compiles on State finances. No focal point to consolidate data at national level. After 73rd and 74th Amendments to the Constitution, local bodies play vital role. Therefore RBI/other organisation should consolidate financial data at national level. CGA compiles monthly data for Central Government and provided to the public through the web-site and State Government data is not available monthly. Hence, there should be a mechanism to consolidate financial data at national level including public sector data. There is still a problem in accrual and cash based accounting principle. CGA and C&AG do place monthly data in their websites. As recommended by 13th finance commission, sub-national data on financial sector would be helpful. Dr. Rakesh Mohan's Committee on financial sector standards could be referred. Further employment and un-employment and labour market data are not available periodically due to conceptual differences.

Representative from CMIE has informed that the user community is looking for integration of data. Also he has informed that building of alternative data sets for students and business community would be very helpful. There is an obligation to user community using latest technology to provide micro-level data.

Representative from NIC informed the committee that NIC is helping Ministry in setting up their computer network and maintaining websites. Using latest technology they serve the ministry for data retrieval and aggregation.

Representative from NSSO informed that NSSO is having an obligation to provide quality data to the user community. Also NSSO supplies micro-level data through Computer Centre. IT tools are deployed to help easy access to data.

DDG computer Centre informed that the data are provided both at macro and micro level. Various reports are placed in PDF format in the Ministry's website. Computer Centre is planning to put up in EXCEL Format also. Data from other Ministries are difficult to make in single format and hence it becomes difficult for generating meta data and exchange of data. At present Computer Centre provides ASI and NSSO data at micro level. Large number of users have complained that they find difficulty in using micro level data. Further, no concordance table is made available for changes between different rounds. Also he informed that there are licence problems in Software and retrieval of data from data warehousing. The members were informed that NSDI can be invited to make presentation on data dissemination standards.

DDG, CSO (IS Wing) has informed that the structure of data management is same as that of NSSO. Involvement of private sector is helpful for value addition.

Chairman informed that the intellectual property right to turn the public data into commercial data should be examined. Electronic standard form should be developed for financial statistics for exchange of data. Further he has stated that the committee should articulate its view on data management which will serve for 10 years ahead and the data gap should be bridged. The tenure of the Committee should be extended for making its usefulness as lot of studies and exercises would be undertaken to assess the best practices available. In this connection he has informed that finalisation of Core statistics is also important before taking up further action. He informed the committee that the Core statistics is Core requirement.

Further, the Chairman desired the committee that the following issues to examine:

- What are the obligation to the user community (both policy makers and others)
- What should be the benchmark, use of machinery and feed back?
- Issue of technology and software
- Issue of building transparency
- Inter-connectivity among the core ministries – Ministry of Information and Technology, Ministry of Finance, RBI and RGI.
- How to produce common single window to provide data for user community?
- How to get user perspective?

Executive Director, RBI informed the committee that core statistics is important prior to data management and other statistical activities. Historical data and time series data are equally important for data analyses and

forecasting. Deployment of technology tools is also important. In this regard RBI can help and make presentation on use of technology tools. Use of GIS, SAS and SDMX should be used in data management. NIC input would be very helpful and is important for building up meta data.

While concluding the meeting, Chairman reiterated the obligation to our people on providing timely, reliable and credible data. There should be a domestic charter for data management – it could be shared once finalised. Core statistics and user community should precede data management. Issues on demand side should be considered through proper interfaces like RBI having interface with CMIE. Corporate sector statistics and partnership with Corporates in producing reliable data should be explored. Geography integration(districts and block level) is key besides integration with planning commission at later stage. During the second meeting, RBI would make presentation on technology being successfully used by RBI.

He desired that few Sub-groups on TOR, Technology, Bench -marking, User community, Intellectual Property, Data Integration Mapping, etc., could be formed for effective specific studies and making inputs to the Committee. The meeting ended with the vote of thanks to the Chair.

Annex

List of Participants:

Sl. NO	Name and Designation	Office
1.	Sh. Suman K. Bery, Chairman,	DG, NCAER, I. P. Estate Member, NSC
2.	Shri Deepak Mohanty, ED, RBI, Mumbai	Reserve Bank of India, Central Office, Mumbai
3.	Sh. K. Kanagasabapathy EPW Research Foundation	EPW Research Foundation, C-212, Akurli Road, Kandivili – East Mumbai -400101
4.	Rep of CMIE	11 Apple Heritage, 54-C, Andheri(E), Mumbai
5.	Sh. P. C. Mohanan, DDG,	Computer Center, MOSPI, New Delhi
6.	Sh. P. C. Sarkar, DDG,	NSSO, DPD, Kolkata
7.	Sh. B. K. Giri, DDG,	CSO (IS Wing), Kolkata
8.	Sh. S. K. Jain, Sr. Technical Director (NIC)	NIC Cell, MOSPI
9.	Sh. M. V. S. Ranganatham, DDG, NSC	NSC, MOSPI
10.	Dr. S. Durai Raju, DDG, CSO (NAD)	NAD, CSO, MOSPI
11.	Sh. G. S. N. Murthy, DDG, NAD	NAD, CSO, MOSPI
12.	Sh. M. Singh	RBI
13.	Sh. M. A. Khan, A. D.	NAD, CSO, MOSPI

(ii) Minutes of the second meeting of the Committee

Annexure-1 b

Second Meeting of the Committee on Data Management – Presentation on RBI's Data Warehouse

1. The Second meeting of the Committee on Data Management set up by the National Statistical Commission was held in Mumbai at Conference Room, Department of Statistics and Information Management, Reserve Bank of India, on February 17, 2011. The meeting was chaired by shri Suman Bery, Chairman of the Committee on Data Management. The meeting was attended by members of the committee as listed in the Annexure. As desired by Shri Deepak Mohanty, Executive Director, RBI and member of the Committee, the DSIM team led by Shri A. B. Chakraborty, Officer-in-Charge made presentation on the RBI's data warehouse covering its functionalities and architecture in the first half the of the meeting.

2. The presentation on data warehouse covered (i) data warehousing concepts in brief also access to Metadata, (ii) overview of the RBI's data warehouse, (iii) live demonstration of the data warehouse covering generation of standard reports, simple queries and advanced queries and (iv) a brief discussion on architecture of the RBI's data warehouse and some issues relating to implementation of a DW project. The members appreciated the wide coverage and various reporting facilities available through the web based interface.

3. During the second half, the Chairman and other Committee members wanted a few clarifications from RBI regarding the policy issues and the process followed during data warehouse implementation and discussed about an action plan on best utilisation of the experiences of RBI in the overall interest of data management of all government departments. The Chairman observed that RBI's Database on Indian Economy / Handbook of Statistics is the first choice of data source for him and indeed many other data analysts.

4. About RBI experience, Chairman asked about the dimension of the efforts RBI put in on data reconciliation, role of specialists vis-à-vis domain knowledge experts, role of outsourcing, scale of manpower, financial resources, priority identification, responsibility allocation, etc. It was informed to the Committee that at the initial phase RBI had the benefit of the experience of Australian Bureau of Statistics, a typical centralized statistical system. In contrast, however, Chairman indicated that in India the statistical system has evolved over a decentralized system and thus data integration poses a serious challenge. Such federal structure obviously has the inherent issues like inter-operability, lack of accepted standards, inconsistency, multiple data sources, inter-agency conflicts, etc. While summarizing the experience of statistical data management systems across government departments and RBI, Chairman drew the attention of the Member Secretary that all these valuable inputs should be suitably placed as Annex in the final report of the Committee.

5. The committee members sought clarifications on (i) choice between proprietary and open source software, (ii) data access policy, (iii) international practices and standards followed, if any and (iv) data quality issues and data integration process. Following clarifications were provided to the committee.

6. **Choice of proprietary and open source software:** The process for setting up of DW system of RBI was started more than 10 years ago. At that time, use of open source software in this area was not so much in vogue. Moreover, while considering the establishment of data warehouse in a large scale, the need for high availability and availability of skilled manpower on the specific tools etc. were taken into account. At the time of choosing the DW tools, the technical advisory committees at various stages took note of these factors. The RBI was guided by the expert advice given by Advisory bodies for choice of the best available software tools. It had been very helpful to include independent academics to offset possible vendor bias. The software chosen was proprietary in nature. However, sufficient care has been taken to ensure that these products are inter-operable.

7. **Data access policy**: The data in the warehouse is "owned" by the concerned departments / organisation units. While allowing access to the data to other users, the data owners' views are respected and accordingly the access permissions are set by them in the data warehouse. The data access policy was formulated on the basis of the need for internal usage / external dissemination on one hand and the advice of the departments concerned. A high level inter-departmental advisory group oversees the data access and other issues relating to the DW.
8. **International practices and standards**: At the time of implementing the RBI data warehouse, the concept of data warehousing in central banking parlance was relatively new. However, at that time, Australian Bureau of Statistics had constructed a data warehouse. Their experience was a useful input to the process that was followed by RBI.
9. **Data quality and data integration issues**: For some of the variables, data collected by different departments have some minor conceptual / coverage differences, resulting apparently divergent numbers. However, differences such as these, and arising out of different versions such as "advance estimates", "provisional estimates", are stored along with the respective data tags. Further as much of the data comes from regulated entities submitting data under statutory / regulatory requirements, the data quality issues faced by the RBI are not very serious.

10. **Metadata Practice:** The Chairman also wanted to know about the issue of appropriate standards (RBI and International Standards) on various aspects of the data in the RBI's DW. He gave the example of the RBI's data on REER in the RBI Bulletin for which the methodology notes were available elsewhere. Shri A B Chakraborty highlighted that RBI, in a special publication called Manual on Financial and Banking Statistics has disseminated to the public about all the primary data information produced within RBI. On the issue of metadata management, it was clarified that the DW contains metadata descriptions.

11. Before proceeding to discuss other issues, the Chairman thanked the RBI team for sharing the data warehouse experience and noted that it will be an important input while framing the roadmap for the national level data warehouse.

12. Subsequent deliberations of the Committee meeting centred around the RBI data warehouse experience and how best this knowledge can be utilised for recommending the development of a data management system spanning across various departments of the government, heterogeneity of data, non-uniform data format, value added services and make available such data through a DW for all users. The Chairman said that global thinking was gradually emerging towards single window concept for transparency in data dissemination to provide access to all. In this context, he referred to three aspects, viz., international best practices, secondary system of checking/auditing data quality and availability of meta data. He also reiterated that the issues of (a) Already published data (b) making data available through single sources and (c) making available micro level data, could be better handled through learning the experiences RBI. These aspects could be documented well in conformity with the global data documentation initiative standard. Further it was clarified that the data management system is for the Government of India not MOSPI alone.

13. Dr. Durai Raju, Member Secretary, drew attention of the Chairman on the specific details of the action points as outlined in the minutes of first meeting of the Committee and presented a brief background of the RBI data management experience covering this. He also indicated that the term of the Committee has since been extended till April 30, 2011. The Chairman said that the work of the Committee has to be expedited. Shri Durai Raju said that based on the deliberations in the meetings held so far, a draft structure of the report will be provided as envisaged by the Chairman.

14. On the remark of the Chairman about the data management issues of government departments, Shri Mohanan, DDG, Computer Centre, MOSPI shared his experience. He highlighted that mainly 3 types of users look for suvey data.

These include: (a) users for aggregate published data, (b) data users looking for multi-disciplinary data, including social sector statistics and (c) micro data users, mainly researchers, emanating from surveys. While micro data are not necessarily provided in users preferred format, sufficient metadata is presented so that users do not face much hardship understanding the structure of the data. However, he mentioned that there are various issues relating to data documentation, ensuring international standards, maintaining userfriendliness of data presentations, etc. The basic problem lies in the framework itself, when data systems across various government departments are not connected, not integrated and not compatible. This has resulted differences in data documentation standard also. In this context, he highlighted the importance of standard like SDMX format – statistical data and metadata exchange. While appreciating RBI effort towards building such enterprise data warehouse, he doubted whether government will take up such massive investments both in terms of money and manpower.

15. Shri S K. Jain, Member, then shared his data management experience in NIC. Particularly he highlighted standards regarding digitization, dissemination (lack of common practice), need for common format of data entry and making it available online, data transmission issues and finally the aggregation problems.

16. On the data compatibility and data aggregation issues, Shri B. Giri, Member shared his experience in detail citing from the NSSO survey rounds, particularly the issue of standards and definitions, heterogeneity of data in employment, education statistics. In order to emphasize the points, he provided dimensional differences in terms of variance arising out of same population, but from varying survey agencies, an investigator using different concepts and definitions and thus reconciliation is always an issue. In addition to the NSSO survey data, he also cited the example of NFHS data. Therefore, in order to overcome such problems at the government level, there is an urgent need to revamp the manpower structure of NSSO. In this connection, he further emphasized that NSC being the highest statistical regulatory body should stressed upon this urgent need of

manpower and lack of uniform definition and the Committee, in particular, make an appropriate recommendation on this. Prof. S. Bose, representative member from ISI Kolkata also pointed to the fact that RBI has done a commendable job and its experience should enrich the development process undertaken by the general governments through this Committee. With regard to Chairman's query on RBI's efforts to put into reconciliation of data, Shri A B Chakraborty clarified that primary data source is banking data which is of uniform standard, not of major problem. However, data cleaning mechanism is in place for external data sources.

17. Shri Sambashiva Rao, Member from IDRBT, spoke of data management issues from a different perspective: people, process and technology. All these should be efficiently managed in so called information lifecycle management. While RBI could get necessary information from the regulated entities, general government may not have such luxury and thus RBI model may not be appropriate for the government departments. As the need and muliti-disciplinary functionalities of government departments vary across data reporting systems, a datamart could be a better solution that an enterprise data warehouse. In addition there is a need to move from 'need to know basis' to 'responsibility to share'.

18. On the other issue of data gap in Municipal and local body finances, Chairman said that the Committee took note of the material provided by RBI. For successful establishment of database system, commitment, identification of core group and support from management/administration are essential.

19. While concluding, the Chairman gave the outline of the nature of the final report coverage and contents. He stressed that the report should contain : (i) nature of recommendations, (ii) priority, (iii) need to know and right to share, (iv) issue of decentralized system, (v) inter operability, (vi) need for assessability and communicability, (vii) need for scanning of published data before conversion into defined data format, (viii) need to accept the external data, (ix) consistency in definition and standards, (x) draw upon insights into problems, (xi) vertical and horizontal issues, (xii) technological and technical issues, and (xiii) training

intervention/ capacity building. Chairman desired that by **20-th March 2011**, draft outline of the final report should be circulated. By **10-th April all comments** should be incorporated before the final meeting. As desired, discussed and decided, the **final meeting** would be held on **18-th April, 2011** before submission of the report.

20. The Chairman thanked the RBI for hosting the meeting and providing valuable inputs. The meeting concluded with vote of thanks to the Chair.

Annex

List of participants

1. Shri Suman Bery, Chairman, Committee on Data Management and Member, NSC
2. Shri B. Sambamurthy, Director, IDRBT
3. Shri S K Jain, Sr Technical Director, NIC
4. Prof. Smarajit Bose, ISI, Kolkatta
5. Shri Ashish Kumar, ADG (NAD)
6. Shri B. K Giri, DDG, CSO(IS Wing)
7. Dr GSN Murthy, DDG(NAD)/ IT Coordinator, MOSPI
8. Shri P C Mohanan, DDG, Computer Center
9. Shri SVR Murthy, DDG(NAD) for Service Sector data
10. Shri M V S Ranganatham, DDG (NSC), MOSPI
11. Dr S Durai Raju, DDG, NAD – Member Secretary

2. The RBI-DSIM team consisted of the following officials

   i. Shri A. B. Chakraborty, Officer-in-Charge
   ii. Dr. Goutam Chatterjee, Adviser
   iii. Shri Vinay Bahuguna, Adviser
   iv. Dr. A. R. Joshi, Director
   v. Dr. Abhiman Das, Director,
   vi. Shri Navas J., Asst. Adviser,
   vii. Shri Sourajyoti Sardar, Research Officer
   viii. Shri A. N. Yadav, Research Officer

(iii) Minutes of the Third meeting of the Committee

1. Annexure-1c

Minutes of the Third Meeting of the Committee on Data Management

……

1. Third meeting of the Committee on Data Management set up by the National Statistical Commission was held on 19th April, 2011 in room no. 205, Sardar Patel Bhavan, New Delhi under the Chairmanship of Shri Suman Bery. The list of participants is annexed.

2. At the outset the Chairman welcomed the Members of the Committee and thanked the Member Secretary for preparing a comprehensive report.

3. Chairman briefed about the motivation of the report and the significance of follow up action. He then requested the Member Secretary to give a brief introduction of the Report.

4. Member Secretary presented the structure of the Report as follows: The subject of data management being technology oriented, its conceptual frame work presented in brief in the first Chapter was explained. Besides various data management disciplines, Statistical Data and Metadata eXchange and Secured Data Centers have been presented in brief so that the essentials of data management get understood. Also the conceptual background which is the foundation on which data management is to be built upon for a big country like India was presented. Status of data sets on key sectors like Education, Labour, Health, Agriculture, Environment and Forest, Trade, Service Sector, Tax Revenue have been discussed and the agencies responsible for collection, compilation and dissemination of data, conceptual framework and methods of data integration and finally availability of statistics, its content, coverage and availability through internet(web-site) have been presented. Then Various challenges posed in official data management have been elaborated. Finally, the recommendations to be taken up immediately and some over a longer period keeping in view the migration of current data schema to technology driven schema to manage the data in Indian Statistical system so as to benefit all types of user communities and to cater to all types of data needs say

micro, macro, aggregates, derived data set, query based and data mining using Data warehousing technology embedded with OLAP/OLAM architecture have been explained.

5. The Chairman further informed the members that there should be a balance between IT and Governance. With powerful IT tools available, Governance can not be sidelined. There should be balance of act between use of IT and Governance. One should not be at the cost of other. Accountability and Responsibility also needs top priority. Centralisation and strategic issues need to be focussed. Centralisation has its own disadvantages. Resource requirement, timeliness and credibility attracts utmost attention. Interface with user communities particularly on free access, user interface, access mode, prioritization, substantive recommendations and User friendly environment should fall within the ambit of the report. After elaborating on these topics, then, he invited the comments of the members on the draft report.

6. RBI, represented by Shri A B Chakraborty appreciated the comprehensive coverage of the report. Particularly, he complimented for the comprehensive coverage of the topics of drawback and challenges. He informed the members that the conceptual framework on data management given in Chapter-1 is very good, brief, coherent and easily comprehensible. However, he stressed that initial condition assessment and study is important for the success of the project. He also asked the Member whether Technology is first or it should be an enabler. He expressed his view that the technology should be an enabler only. Further he asked that accountability and user friendliness should go together. However, the final comment of RBI on the report would be given after return of Shri Mohanty, ED, RBI who is on foreign tour.

7. Member Secretary informed the Committee that the Initial Assessment would be part of the Implementation and Ministry would be taking appropriate steps. Shri A B Chakraborty informed that a Standing Committee or Empowered Committee has to be constituted for going into the minute details of implementation of the data management project at national level as the project is technology driven huge project. Afterwards, the Chairman invited the comments of Dr Kanagasabapathy, Director, EPW Foundation.

8. Dr Kanagasabhapathy informed the Committee that a grand design may not be always successful and a realistic approach should be adopted. He gave the example of Registrar general of India where a massive data has been processed at national level and key results of 2011 Census was released within one month of completion of enumeration. India is having enough of qualified man power and technology power to handle big project like this. He also said that NSC should assume authority. While he opined that data generated should be treated as public goods and the centralised system has its inherent disadvantages.

9. The Chairman intervened and said that the heart of the Issue is Centralisation. He illustrated with an example of World Bank Data and its World Development Indicators (WDI) based on inputs from 200+ countries and maintenance of consistency and comparability. Dr K Kanagasabapathy gave two examples viz. State Government Finance data and GSDP and DDP released by the State Governments and how they lack uniformity, up-to-date and availability. Then the Chairman illustrated the two models of World Bank Data particularly ownership of data and maintaining the consistency globally. Dr. Kanagasabapathy further informed the importance of quality of data, timeliness, coverage and reporting mechanism. Also he has stated that privatization is important and private sector data should be categorised into (a) Financial Sector (b) Social Sector and (c) Service Sector. He informed the Committee that data is disseminated by EPW Foundation in a user friendly manner. The chairman reiterated on quality check and authenticity of data released by private agencies and the issue of intellectual propriety should be taken into account. Then the Chairman asked the other member to view his comments on the draft report.

10. Shri SVR Murthy, NAD informed the committee that to start a big project like this, one should start with Government Administrative data only and scale up the rest in a phased manner. As Technological and Statistical tools should be properly used, sufficient care should be taken. The Chairman intervened to say that the response for Core Statistics is very poor and an incentive measure similar to one proposed by Finance Commission could be thought of. Also he has requested to annex the government dissemination policy notification issued in 1998 and including any revision thereafter. Further, the chairman requested the RBI to explain the

boundary between data availability for decision maker and data availability for users community. Dr Chakaraborty replied that there are standing guidelines based on the recommendations of the high powered Committee on data availability where different levels are determined and incorporated in the system.

11. Shri Sunil Jain, NIC, informed the Committee the importance of use of Technology and implementation right from initial stage of data collection to final stage of data dissemination. At present users are well versed with use of internet and hence the ultimate aim of the Data Management at national level should take care of data dissemination through internet. Also he said that well coordinated approach is the deciding factor for success of a mega project like this should be preceded by a pilot study.

12. Comments on the report were also received by emails. Prof. Bose, ISI, Kolkata appreciated the content and coverage of the report which was in line with the discussions and decisions of the second meeting. Mr Rohit Sabherwal CMIE, made two observations in his comments; one on Data dissemination and the other on acceptance of External data. Member Secretary clarified that both the observations were explicitly dealt in the report. MOSPI will play a coordinating role for compilation and dissemination at national level and data generating agencies (State Governments, other Ministries and Organisations) will continue to disseminate their data. Regarding the acceptance of external data, the report cautioned the use of external data to suitably supplement or complement the existing data sources after ascertaining its credibility. In order to implement the data management project of this dimension at national level, it may be appropriate to create a new division in CSO with dedicated composition of requisite manpower and other resources. It was decided that an appropriate recommendation to that effect would be added in the report. Finally, the Member Secretary outlined the recommendations of the Committee spread over - short term and over a period of time(long term). Also he explained how every stage of data collection to data dissemination would be taken care while implementing the project by a dedicated team.
13. While concluding the Chairman summarised the importance and motivation behind in finalizing the report, highlighting the linkages with Core statistics, significance of COS-Statistics Law being enacted, applications of these recommendations to all forms of government, importance of pilot study, prioritisation based on core

statistics, validation and notification before release and dissemination of data, release of data by authorised agency/persons, decentralised system with strong national level coordinator, Interface with external data, etc. On the issue of identifying data gaps, Chairman felt that this issue should best be left to the data generating agencies or specific institutions/agencies deployed. He finally out lined how to move further on a possible next course of action.

14. The Chairman thanked all the Members for providing valuable comments and inputs. The meeting concluded with vote of thanks to the Chair.

Annex

List of participants

- Shri Suman Bery, Chairman, Committee on Data Management and Member, NSC
- Shri A. B. Chakraborty, Officer-in-Charge
- Dr K Kanagasabapathy, Director, EPW Foundation
- Shri S K Jain, Sr Technical Director, NIC
- Shri SVR Murthy, DDG(NAD) for Service Sector data
- Dr S Durai Raju, DDG, NAD – Member Secretary

ANNEXURE-2

Current national statistical system on key sectors

National Official Statistics:

India is the largest democracy with varying culture every two hundred km and multitude of spoken languages makes the data management system highly complex. But data needs are highly demanding. The Central Statistical Office (CSO) in the Ministry of Statistics and Programme Implementation (MoS&PI) is the nodal agency for a development of the statistical system in the country and coordination of statistical activities among statistical agencies in the Government of India, State Governments and International Agencies like UN, World Bank, IMF, OECD, etc. The collection of statistics for different subject areas, like agriculture, labour, employment, trade, industry, etc. vests with the corresponding administrative ministries. Generally, the statistical information is collected as a by-product of administration or for monitoring the progress of specific programmes/schemes.

Large-scale statistical operations like the Population Census, Annual Survey of Industries, Socio-Economic Survey by NSSO, Economic Census by CSO, Agriculture Census, Livestock Census, etc. are generally centralised, and these cater to the needs of other ministries and departments, as well as State Governments. Similar set up exists in the States. At the apex level is the Directorate of Economics and Statistics (DES), which is formally responsible for the coordination of statistical activities in the State. They bring out statistical abstracts and handbooks of the States, annual economic reviews or surveys, district statistical abstracts, and State budget analysis; work out the estimates of the State Domestic Product and Retail Price Index Numbers and engage in such other statistical activities as is relevant to the State.

Large amount of data exists at the Centre as well as in the States. But they are not integrated either horizontally or vertically. Further getting timely, reliable and credible data on many fields are too difficult and time consuming. And in many cases data are not made available. Before going into the details of various aspects of data management at national level, it is worth to summarize the sectoral availability of data.

1 Agricultural Statistics

In India, the Agricultural Statistics System is very comprehensive and provides data on a wide range of topics such as crop area and production, land use, irrigation, land holdings, agricultural prices and market intelligence, livestock, fisheries, forestry, etc. It has been subjected to review several times so far so as to make it adaptive to contemporary changes in agricultural practices. India has a well-established and internationally known Agricultural Statistical System. It is a decentralized system with the State Governments. State Agricultural Statistics Authorities (SASAs) play a major role in the collection and compilation of Agricultural Statistics at the State level. Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA) at the Centre is the nodal agency for Agricultural statistics at the national level.

Area Statistics:

Statistics of crop area are compiled with the help of the village revenue agency (commonly known as '*patwari*') in the temporarily settled parts of the country and by specially appointed field staff in the permanently settled States under a scheme known as "Establishment of an Agency for Reporting Agricultural Statistics (EARAS)". The States in the North-Eastern Region and two other Union Territories do not have a reporting system, though the States of Tripura and Sikkim (except some minor pockets) are cadastrally surveyed. They compile what are called conventional crop estimates based on personal assessment of the village security-men ("chowkidars"). The three categories of States and Union Territories account for eighty-six, nine and five per cent, respectively of the total reporting area.

In the States that have a *patwari* agency, a complete enumeration of all fields (survey numbers) called *girdawari* is made in every village during each crop season to compile land use, irrigation and crop area statistics. In the States covered by EARAS, the *girdawari* is limited to a random sample of 20 per cent villages of the State, which are selected in such a way that during a period of five years, the entire State is covered.

Crop area statistics of the temporarily settled areas are comprehensive, being based on the complete enumeration method. They are considered fairly reliable because of the *patwari's* intimate knowledge of local agriculture and his ready availability in the village. However, due to an increasing range of functions assigned to the *patwari*, the *girdawari* tended to receive low priority. In order to improve the timeliness and quality of crop area statistics, two schemes are in operation since early seventies namely, the Timely Reporting Scheme (TRS) and the scheme for Improvement of Crop Statistics (ICS).

The TRS has the principal objective of reducing the time lag in making available the area statistics of major crops in addition to providing the sampling frame for selection of crop-growing fields for crop cutting experiments. The TRS sample of villages is also selected in such a way that the entire temporarily settled parts of the country are covered over a period of five years.

Under the ICS scheme, an independent agency of supervisors carries out a physical verification of the *patwari's girdawari* in a sub-sample of the TRS sample villages (in four clusters of five survey numbers each); and makes an assessment of the extent of discrepancies between the supervisor's and *patwari's* crop area entries in the sample clusters. The supervisor also scrutinises the village crop abstract prepared by the *patwari* and checks whether it is free from totaling errors and whether it has been dispatched to the higher authorities by the stipulated time.

CROP PRODUCTION

Estimates of crop production are obtained by multiplying the area under crop and the yield rate. The yield rate estimates are based on scientifically designed crop cutting experiments conducted under the General Crop Estimation Survey (GCES). The GCES covers around 68 crops in 22 States and 4 Union Territories. The number of experiments and their distribution over the strata are made in a manner to be able to obtain the yield rate estimates with a fair degree of precision at the level of the State and each major crop-growing district. The field staff is periodically trained in the conduct of crop cutting experiments.

The Improvement of Crop Statistics (ICS) scheme carries out a quality check on the field operations of GCES under which around 30,000 experiments are supervised by

the ICS staff at the harvesting stage, fifty percent by staff of the Field Operations Division (FOD) of NSSO and the remaining 50 percent by the staff of the SASA. Due to over burden of patwari, quality of data remains the main concern.

CROP FORECASTS

Advance estimates of production for various decisions relating to pricing, distribution, export and import, etc. is required by Government. The Directorate of Economics & Statistics, Ministry of Agriculture (DESMOA) releases advance estimates of crop area and production through periodical forecasts in respect of principal food and non-food crops (food grains, oil seeds, sugarcane, fibres, etc.), which account for nearly 87 per cent of agricultural output. Four forecasts are issued, the first in the middle of September, the second in January, the third towards the end of March and the fourth by the end of May.

These crop area statistics, crop production statistics and crop forecasts are made available in the form of reports and publications and through the Ministry's web-site.

PRODUCTION OF HORTICULTURAL CROPS

There are two main sources that generate statistics of production of horticultural crops. The first is the Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA), which operates a Centrally sponsored scheme "Crop Estimation Survey on Fruits and Vegetables" in 11 States covering 7 fruit and 7 vegetable and spice crops for estimating area and production.

The second source of horticultural statistics is the National Horticultural Board (NHB), which compiles and publishes estimates of area, production and prices of all important fruit and vegetable crops based on reports furnished by the State Directorates of Horticulture and Agriculture. These estimates are apparently based on the informed assessment of local level officials dealing with horticulture and the reports of market arrivals in major wholesale fruit and vegetable markets. Horticulture statistics are made available in the form of reports and publications and through the Ministry's web-site.

IRRIGATION STATISTICS

Irrigation statistics mainly relate to data on area irrigated by different sources and under different crops. The principal sources of irrigation statistics are the crop statistics compiled by the Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA), and the publications of the Ministry of Water Resources. Besides these, some data on irrigated area are available from the administrative reports of State Government departments and the Agricultural Census. Rainfall and weather data are available from the India Meteorological Department (IMD).

AGRICULTURAL PRICES

The Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA) is responsible for the collection, compilation and dissemination of the price data of agricultural commodities. The price data are collected in terms of (a) weekly and daily wholesales prices, (b) retail prices of essential commodities, and (c) farm harvest prices.

Weekly wholesale prices cover 140 agricultural commodities from 620 markets. On receipt of the prices from various State agencies, the Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA) forwards the same to the Economic Adviser, Ministry of Commerce and Industry for monitoring wholesale prices

Retail prices of essential commodities are collected on a weekly basis from 83 market centres in respect of 88 commodities (49 food and 39 non-food) by the staff of the State Market Intelligence Units, State Directorates of Economics and Statistics (DESs) and State Department of Food and Civil Supplies.

Farm Harvest Prices are collected by the field staff of the State revenue departments for 31 commodities at the end of each crop season and published by the DESMOA. It brings out a periodical publication entitled, *Farm Harvest Prices of Principal Crops in India*.

COST OF CULTIVATION OF PRINCIPAL CROPS

To meet the requirement of Minimum Support Price, a comprehensive survey of the Cost of Cultivation of Principal Crops was initiated in 1970-71. The survey is in operation in 16 States and covers 29 crops, the number and choice of crops in each State depending upon their importance to the State.

The Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA) has the overall charge of implementing the survey programme through the Agricultural Universities in 13 States and general universities in three States by providing them cent per cent financial assistance. The Cost of Cultivation Studies are primarily intended for use by the Commission for Agricultural Costs and Prices (CACP). In addition, these data are used by the Central Statistical Office, Planning Commission, other Economic Ministries of Government of India as well as research organisations. The crop input statistics are made available in the form of reports and publications and through the Ministry's web-site.

LIVESTOCK STATISTICS

Data on livestock numbers are collected through a quinquennial Livestock Census that is a complete enumeration of all households with regard to livestock population, poultry, agricultural machinery and fishing craft. The data collected are quite detailed; the livestock is classified according to various species of animals by breed, sex and age. The Livestock Census is a Centrally sponsored scheme coordinated by Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA). The census is conducted by the State Animal Husbandry Departments with the help of their field staff. Reports of the Livestock Census are brought out in two volumes, the first relating to all-India and State-wise data, and the second to the district-wise information. Latest 18-th livestock census were published and made available in the web-site.

LIVESTOCK PRODUCTS

Statistics of Livestock Products are obtained from two sources: (a) annual "Integrated Sample Survey for Estimation of Major Livestock Products", a Centrally sponsored scheme under the Department of Animal Husbandry and Dairying implemented by most of the States; and (b) periodical household enquiries by the NSSO relating to livestock.

The Integrated Sample Survey is a large-scale survey covering 15 per cent of the villages in the country. The survey provides for estimation of livestock numbers as well as major livestock products (milk, meat, wool, eggs and the unit cost of production of milk and eggs).

The NSSO livestock surveys estimate the livestock possessed by the households with details relating to sex, breed, purchase price, market value, disposal of animals, etc.

Further, NSSO consumer expenditure and enterprise surveys include data on household consumption of livestock products and dairy enterprises, respectively.

FISHERIES STATISTICS

Fisheries of India can be broadly classified into two types namely, marine fisheries and inland fisheries. The Fisheries Statistics Section of the Department of Animal Husbandry and Dairying in the Ministry of Agriculture is in charge of compiling the data relating to this sector. At present data on items like fish production, prawn production, fish seed production, disposal of fish catch, preserved and processed items and aquaculture are being collected from State Governments.

A multistage sample survey is used to estimate the fish production from the marine sector through suitable sampling of landing sites of the fishing craft as well as sampling over time of the landings. Data on deep-sea fishing are obtained through reports required to be furnished by trawlers and other deep-sea fishing vessels.

For inland fisheries, the Central Inland Fisheries Research Institute (CIFRI) devised a methodology for collection of data relating to some important still water areas. This involves dividing water sources into two categories namely, fresh water and brackish water bodies each with a distinct ecology, and classifying them further into three groups according to the level of production. Different sampling methods are adopted for assessment of fish production in each group. The fishery statistics are made available in the form of reports and publications and through the Ministry's web-site.

FORESTRY STATISTICS

Reliable forestry statistics are required for planning, policy-making, analysis and decision-making on forestry investment and development programmes. These statistics are collected mainly as a by-product of administrative reports of the State Forest Departments. Besides the FSI, the Indian Council of Forestry Research and Education (ICFRE) is mandated to collect, collate and compile primary and secondary data generated by the State Forest Departments and various Central ministries. The data on the forestry are obtained through a set of periodical reports furnished by the State Forest Departments and other concerned agencies. In addition to details of forest area, the reports provide information on forest products (wood and non-wood), forest land under cultivation, and grazing land, etc.

FSI is using Remote Sensing (RS) technology to collect data on forest cover under three broad classes (dense forest, open forest and mangroves) on a country-wide scale through a biennial survey. Introduction of digital interpretation has helped in reducing the time lag in the availability of the area estimates to just a few months after the completion of the survey.

The Directorate of Economics and Statistics, Ministry of Agriculture (DESMOA) also publishes statistics of area under forests as part of Land Use Statistics according to the definition adopted in the nine-fold classification of land. This includes all land categorised as forests under any legal enactment dealing with the forests or administered as forests whether State or private owned, whether wooded or maintained as potential forest land.

2 Health Statistics

A healthy population is a developmental goal by itself though it is also a necessary ingredient for the other wider goals of social and economic development. In the recent years there have been significant changes in health conditions and the composition of the health sector and simultaneously major transformations have occurred in knowledge and technology, as well as in the political and economic environment. Life expectancy at birth has risen from 32 years at the time of independence to 64 years in 2010. Similarly various health parameters have improved. Although, overall health conditions have improved in India, the current challenges are enormous. India was one of the first countries in the world to intervene in population control as a national programme in 1951.

An efficient Health Information System is a prerequisite for effective administration of health services and achieving the stated goal of "Health for All". Not only health information relating to aspects of health, such as, the existing health condition of the population, morbidity, availability of health facilities, availability of specialists, doctors and other paramedical personnel is essential for this, but demographic data, data on environment and socio-economic variables of the population are also very important for preparing a good health plan and implementing the same. In the context of data requirements for health planning, health-related data for population should provide insights into following areas:

Demographic data: population by age and sex, rural/urban classification, geographical distribution, occupational classification, literacy, religion, marital status, migration, etc.; Vital statistics: birth and death rates, infant mortality rates, life tables, general fertility rates, etc.; Diseases: mortality rates by age and cause of death, morbidity data by age, sex, prevalence of communicable diseases, deliveries and statistics of anti-natal and post-natal care;

Facilities: hospitals, dispensaries, clinics, nursing homes, diagnostic centres, laboratories, equipments – X-ray and other diagnostic equipments, ambulances, beds, etc.; Manpower: doctors, specialists and practitioners in allopathic, homeopathy and other Indian systems of medicine, nurses, pharmacists, lab technicians other supporting staff (their number, qualification, geographical distribution, availability per unit of population); Finance: GNP, Government Revenue and Expenditure, allocation for health, budget estimates, sources of health finance, expenditure on health by voluntary agencies and other NGOs, private expenditure on health, etc.

Health is a State subject and vital and public Health Statistics are traditionally the responsibility of State Health Directorates. The Central Ministry of Health & Family Welfare (H&FW) consists of three Departments namely, Department of Health, Department of Family Welfare and Department of Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy (AYUSH). The Directorate General of Health Services (DGHS) is the technical advisory wing of Ministry of Health & Family Welfare and is responsible for running the various national disease control/eradication programmes. At the national level, Central Bureau of Health Intelligence (CBHI) in the DGHS, Ministry of Health & Family Welfare, dealing with collection, compilation, analysis and dissemination of health data for the country as a whole.

Apart from CBHI, the Rural Health Division of DGHS compiles and publishes *Rural Health Statistics*. This is a six-monthly bulletin, containing information on Government health infrastructure and manpower deployment in the rural areas. This publication also presents data at State and UT level. The National AIDS Control Organisation (NACO) under Department of Health collects data on cases and deaths due to AIDS/STD and publishes these in its *Annual Update*. CBHI also takes the data on important items from the Rural Health Division and NACO as well as from Department of Family Welfare and Department of AYUSH and publishes the same in *Health Information of India*.

The Department of Family Welfare is responsible for implementing programmes for population control and maternal and child health now renamed as Reproductive and Child Health. The Family Welfare programme is a Centrally- sponsored programme implemented by the respective States and UTs. The information flow starts from the peripheral level where the service delivery takes place. Department of Family Welfare publishes *Family Welfare Programme in India Year Book,* annually.

The Department of AYUSH propagating Ayurveda, Unani, Sidha, Homeopathy, Yoga and Naturopathy. It collects information related to these areas and publishes these in: (a) *Indian Systems of Medicine & Homeopathy in India* (Annual), (b) *Ayurvedic & Siddha Medical Colleges in India* (Quinquennial), (c) *Homeopathic Medical Colleges in India* (Quinquennial) and *Unani Medical Colleges in India* (Quinquennial). In addition, Department of AYUSH also obtains information through surveys on "Usage and acceptability of AYUSH," and "Demand assessment of Medicinal Plants".

Data on medical and health infrastructure (education and treatment) and manpower information are generated as a by-product of administrative and regulatory procedures. The licence registers for various categories of doctors, dentists, pharmacists, nurses, health visitors, etc. provide data about manpower and are consolidated by statutory councils such as the Medical Council of India, Dental Council of India, Nursing Council, etc.

3 Education Statistics

Education is the key to all processes of human development. As such, educational planning needs to be done meticulously and executed with great sensitivity. Education improves the quality of life and develops manpower for different sections of the economy. It empowers the poor masses to become self-reliant and enables them to participate in the process of national development. Education is a concurrent subject, which implies a meaningful partnership between the Union Government and the States. Two departments concerned with education in the Ministry of Human Resource Development, one for Elementary Education and Literacy and the other for Secondary and Higher Education, closely interact with the States and UTs. This task requires the support of a robust mechanism for collection, compilation, processing and dissemination of Educational Statistics. Efficient decision-making has to be based upon a large amount of information and quantitative data. While the policy makers and administrators experience the need for comprehensive database for policy formulation and monitoring of programmes,

the researchers and other data users also feel the need for updated, reliable and inter-temporally comparable data and information.

The Educational Statistics System in India dates back to the pre-independence period. To assess the status and to prepare a plan to this effect, the HRD Ministry conducts the All-India Educational Survey (AIES). These surveys have become an integral part of the system of Educational Statistics in India.
The two main sources of educational data are the educational institutions and households. The educational institutions provide the data on enrolment and number of teachers, which is collected annually from all recognised institutions being compiled at the national level by the Planning, Monitoring and Statistics Division (PMSD) in the Department of Secondary and Higher Education (DS&HE) of the Ministry of Human Resource Development (MHRD). More detailed statistics on students, teachers and physical facilities in schools up to higher secondary level are collected in 5 to 7 years through All India Educational Surveys (AIES) conducted by the NCERT. Important educational data that can be collected only from the households relate to such items as literacy and the educational level of the population, whether the person or child is attending school or not, and private expenditure on education. Data on literacy, level of education and schooling status are collected in the decennial Population Census. The data on these and some other items such as expenditure on education and school dropouts are also collected in certain rounds of National Sample Surveys.

There are various agencies involved in the collection of data on technical and higher education in the country. This area comprises higher (general education), technical education, medical education, agricultural education and teacher education. The University Grants Commission(UGC) is responsible for collection and reporting of data on higher education obtained directly from colleges and universities. The Central Bureau of Health Intelligence (CBHI) in the Department of Health collects data on education in allopathic and dental systems of medicine from the Medical Council of India (MCI) and the Dental Council of India (DCI). The Department of AYUSH collects and publishes data on ayurveda, unani, siddha and homeopathy education.

Mention may be made of the statistics collected annually from schools for the primary level of education in the districts covered under the District Primary Education Programme (DPEP). The DPEP is a Centrally sponsored scheme providing a special thrust to achieve universalisation of primary education through decentralised management, participatory processes, empowerment and capacity building at all levels.

4 Labour Statistics

The subject 'Labour" is a in the Concurrent List of the Constitution, both the Union and State Governments have powers to legislate on issues concerning Labour; their conditions of work, welfare, safety, health, etc. The main agency involved in the collection and compilation of Labour Statistics mainly in the organised sector is the Labour Bureau in the Ministry of Labour. The Labour Bureau collects statistics through statutory and voluntary returns under different Labour Acts. The State

Governments compile such statistics at the State level; the Bureau in turn consolidates them for the country as a whole covering all States and sectors of the economy and brings out periodical reports. It also conducts occasional surveys concerning labour in specific geographic areas or for some specific section of labour. Essentially, these are either to study the socio-economic conditions of labour with a view to formulating policy measures or to assess the impact of labour enactments.
The Labour Bureau in the Union Ministry of Labour has been the Central agency collecting and disseminating data on various aspects of labour. In Labour Statistics the involvement of the State Governments is crucial for the improvement of the system. The major agencies involved in the collection of Labour and Employment Statistics are the Ministry of Labour and its affiliates such as Labour Bureau and the Director General of Employment and Training (DGE&T); the National Sample Survey Organisation (NSSO); and the Registrar General and Census Commissioner of India. The Central Statistical Organisation (CSO) also collects data on employment through the Economic Census.

The DGE&T brings out a number of publications based on the data collected through the Employment Market Information Programme (EMIP) and the National Employment Service (NES). The data collected through EMIP is disseminated through various publications, which provide estimates of the utilisation of labour force in different sectors, industries, occupations, etc. the excess and shortage of manpower and the present level of employment generation in various industries.

5 Industrial Statistics (Organized sector)

For organized sector, Annual Survey of Industries (ASI) conducted by Ministry of Statistics and Programme Implementation is the major formal source for a comprehensive database on the different aspects of industrial statistics in the country. A number of measures, particularly those taken in the recent past, for speedy collection of primary data, quick data processing and data dissemination have led to the development of a system where the time lag in releasing the ASI results has been reduced considerably.
The Annual Survey of Industries is the principal source of Industrial Statistics in India. It plays a key role in assessing the changes in the growth and structure of the registered units in the manufacturing sector. With the implementation of the Factories Act, 1948, the coverage was extended to all factories employing 10 or more workers and using power, or 20 or more workers but not using power on any day of the preceding 12 months.
The ASI covers the following categories of units:
All factories registered under sections 2m (i) and 2m (ii) of the Factories Act, 1948 employing 10 or more workers and using power, or 20 or more workers but not using power on any day of the preceding 12 months;
All *bidi* and cigar manufacturing establishments registered under the *Bidi* and Cigar Workers (Condition of Employment) Act, 1966, employing 10 or more workers using power, or 20 or more workers without using power;
Units engaged in certain services, repair of motor vehicles and a few other consumer durables like watches, etc. employing 10 or more workers using power, or 20 or more workers without using power; and

The units or factories in the ASI frame are grouped into census and sample sectors. While the factories in the census sector are surveyed on a complete enumeration basis, a representative sample from the sample sector is considered for survey in any survey year. For unorganized sector, data are generated through various surveys.

6 Foreign Trade Statistics

The Directorate General of Commercial Intelligence and Statistics (DGCI&S), in Ministry of Commerce and Industry is the nodal agency for collection, compilation, publication and dissemination of Foreign Trade Statistics. The main sources for India's Foreign Trade Statistics are Shipping Bills and Bills of Entry – declarations made and submitted by exporters and importers, respectively to the authorities of customs at the ports. These bills are statutory documents, which contain the customs' permission to ship or land the goods, as the case may be. These Shipping Bills and Bills of Entries for each item of export and import contain relevant details of the transactions such as Code Number of the Commodity according to Indian Trade Classification based on Harmonised Commodity Description and Coding System {ITC(HS)}, description of the commodity, license particulars of the goods in the case of imports, value of exports or imports, quantity (gross and net), amount of duty, name of exporter or importer, country of destination or consignment, Importer and Exporter Code (IEC), etc.
Non-commercial transactions such as, personal baggage and effects, exhibition goods, samples, etc. are not covered in Foreign Trade Statistics. The direct transit trade, i.e. goods of other countries passing and transit goods warehoused not for the purpose of disposal, are excluded completely. The exports and imports of crude oil and petroleum products are included in the Foreign Trade Statistics.

The DGCI&S receives trade data from about 40-50 major ports and some small ports in three different modes namely, Electronic Data Interchange (EDI), Non-EDI and manual. The data transcribed manually from Shipping Bills and Bills of Entry into DTR formats has about a 15 per cent share in the total merchandise trade, the remaining 85 per cent share being accounted for by EDI and non-EDI modes. In the EDI mode, customs clearance of Shipping Bills and Bills of Entry are given through the computer itself and the relevant information as required in the DTRs is furnished to DGCI&S. The major ports are covered by the EDI system. The non-EDI mode relates to manual clearance of Shipping Bills and Bills of Entry but the DTRs are prepared through the help of the computer and the data are transmitted to the DGCI&S in floppies.
The DGCI&S prepares the aggregated data of Exports and Imports within about 25 days from the close of the month in the form of a Draft Press Note. As Quick Estimates of Exports and Imports by Principal Commodities are prepared at the same time with a view to make a broad assessment of the performance of India's Foreign Trade and to study the impact of various trade policies, enabling the Government to initiate suitable corrective measures, if necessary, in attaining the desired objective in the sphere of External Trade. A monthly brochure entitled, *Foreign Trade Statistics of India (Principal Commodities and Countries*) is then brought out in about two months time from the reference month. Detailed data of India's Foreign Trade are released in two publications, namely, *Monthly Statistics of*

Foreign Trade of India (MSFTI) containing Commodity by country details and *Statistics of Foreign Trade of India by Countries (SFTIC)* containing country by commodity details. The data reported in these publications are according to 8-digit ITC (HS) and is available with a time lag of about 3 to 4 months from the reference month.

7 Service Sector Statistics

The Services Sector constitutes a large part of the Indian economy both in terms of employment potential and its contribution to national income. The Sector covers a wide range of activities from the most sophisticated in the field of Information and Communication Technology to simple services pursued by the informal sector workers, for example, vegetable sellers, hawkers, rickshaw pullers, etc. Activities Comprising the Services Sector are Trade, Hotels and restaurants, Transport including tourist assistance activities as well as activities of travel agencies and tour operators, Storage and communication, Banking and insurance (banking statistics dealt separately in this section), Real estate and ownership of dwellings, Business services including accounting; software development; data processing services; business and management consultancy; architectural, engineering and other technical consultancy; advertisement and other business services, Public administration and defense, Other services including education, medical and health, religious and other community services, legal services, recreation and entertainment services, Personal services and activities of extra-territorial organisations and bodies.

Although the Services Sector has a very pivotal role in the country's economic development, the database in this Sector is highly disorganised. A major limitation of the existing statistical system in this respect is the absence of a well-organised mechanism for maintaining a regular and proper database for this Sector. There is no scheme in the Services Sector for annual collection of data from the units either having a large number of workers or contributing significantly in terms of annual turnover. The main difficulty in this regard is the non-availability of an up-to-date frame of such units. The problem of data collection from this Sector through the Follow-up Enterprise Surveys of Economic Census is compounded by the fact that the Sector is dominated by a large number of *unorganised* units. Further, the composition of units in the domain undergoes changes at a rapid pace because new units or newer service areas come into existence and others disappear with alarming frequency. Thus, a sound official statistical system should endeavour to address all these methodological issues for properly estimating the size and contribution of the Services Sector marked by a rapid change in its composition.

The Services Sector of the economy can be broadly grouped into three broad segments namely, the public sector, private corporate sector and the household sector. The first two are generally referred to as the organised part of the economy, as the accounts of all the business transactions of these sectors are recorded in specified documents and are made available as public documents at regular intervals. The remaining part of the economy, that is the household or unorganised sector, constitutes all unincorporated enterprises including all kinds of proprietorship and partnerships run by the individuals.

As regards the organised sector, the data contained in various budget documents or reports and the accounts provide the basis for the estimates for the public sector.

For the Private Corporate Sector, the annual reports of the companies are the main source of data for estimation of number of workers, gross value added (GVA), etc. The estimation of various characteristics is based on the company finance studies carried out by the Reserve Bank of India taking the annual reports of a sample of companies. The estimates so obtained are suitably inflated to derive the current estimates for the entire population of joint stock companies registered with the Registrars of Companies (ROCs) in the various zones or States of the country.

The Follow-up Enterprise Surveys of the Economic Census (EC) periodically conducted by the Ministry of Statistics and Programme Implementation (MoS&PI) provide estimates of number of enterprises, workers, GVA, etc. for the unorganised Services Sector. The estimates of GVA per worker obtained from these surveys are used as a benchmark for the subsequent years, till the results of the next survey on the same subject are available. The benchmark estimates of GVA per worker are carried forward using suitable indices.

As regards data relating to the unorganised Services Sector, the estimates of GVA per worker based on the Follow-up Enterprise Surveys of EC. Further, the estimates of the number of workers in different sub-sectors as per these surveys also available from other sources like Employment-Unemployment Surveys of the National Sample Survey Organisation (NSSO) and decennial Population Censuses.

The departments outside the MoS&PI generate a huge volume of data, either as a by-product of their regular activity or through studies meant for generating the required data. With the increasing use of computers and telecommunication in business transactions, the domain of the Services Sector is growing bigger day by day. The software industry, and particularly its share in external transactions, has grown at a rapid pace during the last decade. Much of these external transactions are carried through the Internet, and thus are not reflected in the regular imports and exports data obtained from the custom ports. NASSCOM provides annual data on e-commerce.

8. Banking Statistics:

Banking system in India comprises commercial banks and cooperative banks with a vast branch network across the country to cater to the financial transaction needs of the people. The commercial banks are broadly categorized into public sector banks, domestic private banks and foreign banks. The public sector banks comprise State Bank of India and the associate banks and the nationalized banks as also the regional rural banks. While the banks themselves primarily manage the huge amount of information generated as a result of the financial transactions for their own requirement, the Reserve Bank of India (RBI), in course of performing central banking functions and keeping in view the overall economic perspective of the banking system, collects, compiles and disseminates a host of banking and monetary statistics in India.

Under the Reserve Bank of India Act, 1934, RBI has been entrusted to perform a wide range of functions such as monetary management, currency management, financial regulation and supervision, foreign exchange and reserves management, government debt management, acting as banker to the banks and to the Government and an active developmental role, particularly for the agriculture and rural sectors. In addition, under the Banking Regulation Act, 1949, RBI has the

responsibility to regulate and supervise banks' activities in India and their branches abroad. Besides, the National Bank for Agriculture and Rural Development (NABARD) also compiles a variety of banking statistics, especially in the rural and cooperative sector.

RBI has put in place an elaborate mechanism for management of the vast amount of banking data collected through various statutory and non-statutory returns. The information collected includes assets and liabilities of banks, spatial distribution of deposits and credit, sectoral distribution of credit, interest rates, international banking business, priority sector advances, etc. The statistical systems developed for managing these data include prescription of reporting systems; monitoring of information flow; computerized systems for processing and generation of a variety of reports; and dissemination of information in various print and electronic forms. Against this background, a brief account of the data management systems for banking information is presented below.

9. Socio-Economic Statistics

Timely and reliable Social Sector Statistics is vital for the effective development of social policy, informed decision-making and for evaluation of the impact of social and economic policies. Incomplete and inadequate collection and compilation mechanism of Social Statistics for the planners and policy makers can therefore constitute a major impediment to effective social development of the country. The importance of the linkages between policy development and Social Statistics points to the need for greater national priority to be given to Social Statistics.

Socio-economic Statistics thus form an important component in the development of the country and include a vast array of information on health and disease; literacy and education; standard of living and poverty; labour force and employment; status of women and gender empowerment; population parameters relevant to fertility, mortality and migration; ecology and environmental protection. Timely collection of appropriate, adequate and reliable data on the above dimensions and proper use of this in planning, implementation, monitoring, evaluation and redesign of various developmental programmes/schemes is absolutely essential if the country has to progress more rapidly and join the ranks of the developed economy.

In India, the concerned ministries and departments of the Union Government are engaged in the collection, compilation and dissemination of Socio-economic Statistics through the corresponding departments in the State Governments in prescribed formats. Many of the data series are a by-product of the general administration of the States based on the records of the concerned offices, as also a product of the administration of particular Acts of the Government and Rules framed under them. This system generates data on a wide range of subjects in the Social Sector.
Increasing requirement and demand is being felt for decentralised databases on population size along with its characteristics for purposes of micro level planning in various development programmes initiated following the democratic decentralisation process set in motion by the 73rd and 74th Constitutional amendments that gave greater responsibilities and powers to the *panchayats* and nagar palikas. Therefore, the thrust of the recommendations is on improving the existing mechanism of

administrative data collection with the trust and responsibility largely placed in the concerned agencies to compile needed data at as disaggregated a level as possible.
As much of the responsibility for producing timely, credible and adequate statistics lies with the administrative Ministry and Departments both at the Centre and in the States, the powers and responsibilities vested in it for data collection, analysis and dissemination. Ministry of Health and Family Welfare, Ministry of Rural Development, Ministry of HRD, Ministry of Social Justice and empowerment, Ministry of Panchayat Raj at the Centre publishes the data for their subject domain and disseminate by web-site as well. NSSO conducts nation-wide Surveys on various aspects of Socio-economic characteristics of the population and publishes them through periodical reports and publications besides making available through Ministry's web-site.

10. Revenue Statistics

The Budget documents of Centre and State Governments provide details of taxes, both direct and indirect taxes. The authority to levy taxes is divided between the Union Government and the State Governments under the relevant Acts. The Union Government levies direct taxes such as personal income tax and corporate tax, and indirect taxes like custom duties, excise duties and central sales tax. The States are empowered to levy State sales tax and other local taxes like entry tax, octroi, etc.
The Department of Revenue, Ministry of Finance is responsible for all matters relating to the administration of Central taxes. The Central Board of Direct Taxes (CBDT) administers the direct taxes through its subordinate organisation namely, Income Tax Department while the Central Board of Excise and Customs (CBEC) is responsible for the administration of indirect taxes through Departments of Customs and Central Excise.
The Research and Statistics Wing of the Directorate of Income Tax (RSP&PR) of the CBDT is engaged in the collection and compilation of direct tax statistics. The data flow from the field offices of the Commissioners and Chief Commissioners of income tax to the Directorate of Income Tax (RSP&PR), where they are consolidated at the all-India level. The Directorate prepares various statistical statements and reports of different periodicities (monthly, quarterly and annual) based on the information received from the field offices. These reports and statements, essentially meant for departmental use, cater to the needs of the CBDT and Ministry of Finance.
The CBEC is responsible for levying, collecting and monitoring of Central excise and custom duties all over the country. The data pertaining to Central excise and customs are collected under Central Excise and Custom Law and Rules framed thereunder. Statistics relating to Central excise are generated on the basis of statutory return filed monthly by each Central excise assessee. The data flow from the range office (lowest formation of Central excise) to Commissionerate office through divisional offices and finally, to the Directorate of Statistics and Intelligence, which in turn submit the Reports to CBEC.

11. ENVIRONMENT STATISTICS

Ministry of Environment and Forests is the nodal agency for maintenance of proper statistical system related to environment. Environmental issues are enshrined in the Indian Constitution as Directive Principles of State Policy and reflect the commitment of the country to protect the environment with regard to forests and wildlife. In

India, Ministry of Environment and Forests is engaged in the task of managing the country's environment by focusing on the development of important tools and techniques, impact assessment, research and collection and dissemination of environmental information.

One of the important challenges for developing economies like India, is to achieve closer integration between economic policies and policies for management of natural resources and environment. For closer integration, decision-makers need more information about the environmental impacts of developmental policies and identification of pressure points such as population, industrialisation, large power generation and irrigation projects, etc. This calls for collection, compilation and dissemination of a wide variety of statistical data on biodiversity, atmosphere, land and soil, water, human settlements, etc. Environmental Statistics, which has become an important area requiring special attention, has three major sub-areas, namely, basic Environmental Statistics, environmental indicators and environmental accounting.

An Environmental Statistics Cell has also been established in the CSO, to coordinate with various agencies involved in collecting information relating to Environment Statistics. The State Governments have also set up departments dealing with Environment to address the rapidly increasing policy initiatives and programmes in the environment and forest sectors.

Apart from the CSO, various ministries and departments of Central and State Governments collect information related to Environment Statistics and the same are published in various publications namely, *Forestry Statistics, The State of Forest Report, Inventory of Forest Resources of India, State of Environment,* etc. by organisations within the Ministry of Environment and Forests; *Agriculture Statistics at a Glance* and *Fisheries Statistics* by the Ministry of Agriculture; *Water Statistics* by Ministry of Water Resources, etc. Most of these publications are annual, but the time lag in bringing out the publications is about 3 to 5 years. Information on some indicators is also being collected by other agencies like the Central Pollution Control Board, Registrar General, Ministry of Urban Development, Tata Energy Research Institute, etc. The Ministry of Environment and Forests has established an Environmental Information System (ENVIS) for maintaining information on various aspects related to environment. As environment is a multi-disciplinary subject involving complex subjects like bio-diversity, atmosphere, water, land and soil and human settlement, it is very difficult to collect and analyse data and study inter-relationships. As there are several agencies for collection of information, it becomes necessary to develop an efficient statistical system on environment that could meet the growing demand of various stakeholders, both Government and non-Government as well as outside agencies for environmental data.

12 Price Statistics

INTRODUCTION

The price indices are closely watched indicators of macro-economic performances. They are direct indicators of the purchasing power of money in various types of transactions involving goods and services. As such, they are also used as deflators in providing summary measures of the volume of goods and services produced and consumed. In India, for these varied purposes Central and State Government agencies collect the primary data on prices. There are mainly three agencies

namely, Labour Bureau in the Ministry of Labour, Office of Economic Adviser in Ministry of Industry and Central Statistical Office in the Ministry of Statistics and Programme Implementation responsible for the compilation and release of various indices. This section deals with the issues relating to a range of consumer price indices, wholesale price index number and the price collection mechanism.

CONSUMER PRICE INDEX NUMBERS

At the national level, there are four Consumer Price Index (CPI) numbers and one Wholesale Price index. These are:

- **A.** CPI for Industrial Workers (IW),
- **B.** CPI for Agricultural Labourers (AL),
- **C.** CPI for Rural Labourers (RL) and
- **D.** New CPI for Rural and Urban and
- **E.** Wholesale Price Index.

The base years of the current series of CPI(IW), CPI(AL) and CPI(RL), and CPI(New Series) are 1982, 1986-87, 1986-87 and 2010, respectively. While the first three are compiled and released by the Labour Bureau in the Ministry of Labour, the fourth one is released by the Central Statistical Office in the Ministry of Statistics and Programme Implementation.

CPI for Industrial Workers, CPI(IW)

The Current series of CPI(IW) on base 1982=100 covers industrial workers employed in any one of the seven sectors namely factories, mines, plantation, railways, public motor transport undertakings, electricity generation and distribution establishments as well as ports and docks. The index covers only manual workers irrespective of their income.

The total number of centres (70) was allocated among the factory, mining and plantation sectors in proportion to the total employment in the country. Though, a centre is selected on the basis of workers in 3 sectors namely Factory, Mining and Plantation, but the detailed survey covered all the 7 sectors.

Centre-wise, the item basket was determined on the basis of WCFIES. The items retained in the basket accounted for a substantial share of expenditure in the group or sub-group of items and could also be allotted a price over the life of the series. The all-India index is a weighted average of 70 centres' indices. The weight assigned to each centre is the proportion of the estimated consumer expenditure of the centre to the aggregate consumer expenditure of all the centres.

CPI for Agricultural Labourers and Rural Labourers, CPI (AL/RL)

A person is treated as an agricultural labourer if he or she follows one or more of the agricultural occupations in the capacity of a labourer on hire, whether paid in cash or kind or partly in cash and partly in kind. A rural labourer is defined as one who does manual work in rural areas in agricultural and non-agricultural occupations in return for wages in cash or kind, or partly in cash and partly in kind.

For the purpose of collection of consumer expenditure data for deriving weighting diagrams for CPI(AL/RL) as a part of general consumer expenditure survey of NSSO, the rural labour household is defined as one which derives its major income during the last 365 days from wage paid manual employment (rural labour), *vis-à-vis* wage paid non-manual employment as also self-employment. From amongst the rural labour households, those households which earn 50% or more of their total income

(from gainful occupation) during the last 365 days from wage paid manual labour in agriculture are categorised as Agricultural Labour Households. This series is presently compiled for 20 States and All-India. Monthly price data collected from 600 villages spread over 20 States by the field staff of FOD are used in the compilation of these indices. The sample of 600 villages is staggered over four weeks of a month with one-fourth of the sample covered every week. Prices are collected on the fixed price collection day which may be a "Hat" day for "Hat" or non-daily markets and any market day for daily markets. CPI(AL) is a sub-set of CPI(RL) series. The rural retail prices for these two index series are the same but the weighting diagrams are different.

Consumer Price Index (new Series)

Consumer Price Indices (CPI) measure changes over time in general level of prices of goods and services that households acquire for the purpose of consumption. CPI numbers are widely used as a macroeconomic indicator of inflation, as a tool by governments and central banks for inflation targeting and for monitoring price stability, and as deflators in the national accounts. CPI is also used for indexing dearness allowance to employees for increase in prices. CPI is therefore considered as one of the most important economic indicators.

CPI numbers presently compiled and released at national level reflect the fluctuations in retail prices pertaining to specific segments of population in the country viz. industrial workers, agricultural labourers and rural labourers . These indices do not encompass all the segments of the population in the country and as such do not reflect true picture of the price behavior in the country. To overcome the above, the Central Statistics Office (CSO) of the Ministry of Statistics and Programme Implementation compiles new series of CPI for the entire urban population, viz. CPI (Urban), and CPI for the entire rural population, viz. CPI (Rural), which would reflect the changes in the price levels of various goods and services consumed by the urban and rural population. These new indices are compiled at State/UT and all-India levels.

Compilation of CPI numbers generally consists of two stages i.e. (i) calculation of price indices for elementary aggregates (item level indices) and (ii) the aggregation of these elementary price indices to higher level indices using the weights associated with each level. Laspeyre's formula is used for aggregation of indices. Specifications of items have been selected on the basis of popularity in the respective areas. These specifications are different in terms of units, quality etc for different price schedules. Prices relative of each product specification (current month price/base year average price) is worked out. Average of these price relatives under the respective item multiplied with 100 gives the index for that item.

Index Numbers for both rural and urban areas and also combined are released w.e.f. January 2011. Provisional indices based on the data available are first released with the time lag of 18 days. State/UT indices are released only if receipt of schedules is 80% or more of the total schedules allocated to a state/UT. These provisional indices are subsequently revised and final numbers with complete data for all India and also for all the States/UTs are released with a time lag of two months. Indices

for January 2012 onwards, along with annual inflation rates would be released with a time lag of one month.

The Wholesale Price Index (WPI)

The Wholesale Price Index (WPI) series with base 2004-05 is compiled by the Office of Economic Adviser (OEA), Ministry of Industry, on a weekly basis, based on the price quotations collected by the official as well as non-official source.

The commonly-used measure of inflation in the Indian economy is based on the WPI. As WPI measures the price change at the level of either the wholesaler or the producer and does not take into account retail margins, it thus represents the production side and not the consumption side. For a true measure of inflation, it is necessary to measure the changes in the prices only at the final stage of transaction.

13 CORPORATE SECTOR STATISTICS

The Ministry of Corporate Affairs (MCA) is the nodal agency for collection, compilation and dissemination of Corporate Statistics. The growing importance of the Corporate Sector calls for greater transparency and availability of data. Furthermore, the withdrawal of direct regulatory functions by the Government such as industrial licensing, import licensing, capital issues and exchange controls means that a number of avenues of collection of data have ceased to exist while the need for them has grown for indicative planning, forecasting and research purposes. Finally, the onset of the knowledge-based sectors or the new economy requires better reporting standards of certain attributes to help monitor the national economic performance and to assess its future prospects.

The MCA has recently introduced a scheme of assigning a unique 21-digit Corporate Index Number (CIN) for registration of companies. The CIN has been designed to help easily identify or group the companies by State, industry (whether listed or not), economic activity, ownership and year of incorporation and will be applicable to all companies registering themselves from 1 November, 2000. The older companies will also be given the new registration number subsequently. All the Registrars of Companies (ROCs) will be brought under a network to facilitate the monitoring of the submission of various documents under the Companies Act. This facility will enable the identification of defunct companies, once the complete database is prepared. As computerisation is being introduced in a phased manner, the availability of the database is likely to take some time.

14. National Accounts Statistics

The national accounts provide a comprehensive, conceptual and accounting framework for analyzing and evaluating the performance of an economy. CSO is mandated to compile and disseminate National accounts statistics. CSO releases quarterly GDP, Advance Estimates of GDP and Quick Estimates of GDP as per the release calendar.

Advance estimate of National income and its update

The first estimates of the Annual National Income for a reference year are released by the CSO, about two months before the close of the year, in the form of Advance

Estimates (AE) of National Income. These estimates present at both current and constant prices and at factor cost, the Gross National Product (GNP), Net National Product (NNP), Gross Domestic Product (GDP), Net Domestic Product (NDP), and Per Capita Income (Per Capita Net National Product at factor cost) by industry. These estimates are subsequently revised and released on the last working day of June, i.e. with a lag of three months, as updates of advance estimates. The AE are compiled using the methodology evolved by the CSO which is similar to the methodology adopted for the Quick Estimates (QE) and are based on anticipated agricultural production and industrial production, analysis of budget estimates of Government expenditure and performance of key sectors like railways, communication, banking and insurance, available at that point of time.

Quick Estimates of National Income and related aggregates

Quick Estimates of NAS and the Revised Estimates of the earlier years are released by the CSO utilising the available data of various sectors provided by the statistical system, in the month of January or February of the following year (with a 10-month lag). Along with the Quick Estimates for the previous financial year, estimates for the earlier years are also revised using the detailed data supplied by various source agencies.

Quarterly Estimates of GDP:

The Central Statistical Office (CSO) has introduced the quarterly estimates of Gross Domestic Product (GDP), w.e.f., the estimates relating to January-March 1999, both at constant (1993-94) and current prices, on 30.6.1999. The quarterly GDP estimates as per advance release calendar will henceforth be released by the CSO on the last working day of a month preceding each quarter, the estimates referring to the previous quarter. The introduction of quarterly GDP estimates into the statistical system of the country is a result of India's subscribing to the Special Data Dissemination Standards (SDDS), introduced by the International Monetary Fund (IMF). At present, Quarterly GDP are released both supply and demand side.

The quarterly GDP estimates take into account the quarterly performance of various sectors of the Indian economy. The methodology of preparing quarterly estimates involves three steps, namely, (i) preparing quarterly estimates at constant (2004-05) prices for the benchmark year (The benchmark year is dynamic and refers to the latest year for which fully revised annual GDP estimates and quarterly data on the indicators are available. For the first set of quarterly estimates now released by the CSO, the benchmark estimates refer to the year 1996-97), (ii) preparing quarterly estimates for the reference quarter at constant (2004-05) prices, and (iii) estimating the implicit price deflators for the reference quarter to derive the quarterly estimates at current prices.

15. Integration of Official Statistics

The Integration of Survey and Administrative data represents a first attempt to create a common methodological basis for the application of statistical methodologies for the integration of different sources. It aimed at reviewing and promoting knowledge and application of sound methodologies for the joint use of available information in existing data sources for the production of official statistics by putting together available data in a user friendly manner.

The advantages of having an 'integrated and coordinated approach' to the production of official statistics have been recognized for a long time. One result has been that the national statistical system is based on a model where there is a strong central statistical office (MOSPI) responsible for producing statistics for a wide range of areas and users, and for maintaining the implementing establishment and household surveys. Frequently specialized units in line ministries and operational agencies will supplement the MOSPI, with the task to produce statistics on the basis of administrative records that are created for operational reasons by the same agencies, and also to analyze the resulting statistics together with statistics produced by the MOSPI, as a basis for formulating, implementing and evaluating policies for which the agencies are responsible.

The precise division of business and areas of responsibility between the MOSPI and other line Ministries have been identified through allocation of business rules by government of India. The line Ministries will, however, often mirror to some extent the former as the policy formulating and executing agencies are important users of the statistics produced by the MOSPI: thus we get units for statistics on health, education, employment and unemployment, wages and income, social security, population and families, as well as on prices, wholesale and retail trade, foreign trade, transport, manufacturing, construction and public finance, to mention just some of the possible 'areas of statistics' that may be identified in the allocation of business rules of Government of India. Those responsible for official statistics have also recognized for a long time that the 'real world' these statistics should aim to describe cannot be divided neatly along lines that correspond to those which are convenient for organizing the production of official statistics: because in the final count "everything depends on everything else".

One can get inspired by the power and success of the *National Accounts Statistics(NAS)* based on SNA accounting framework as a coordinating tool for different areas of 'economic' statistics in the country. Similar efforts to create similar tools for health, education, social sector statistics, environmental statistics, labour statistics, demographic statistics etc. required. Integration at subject level by the concerned Ministry at national level is most important. As an implementation agency of the various policies at national level they are better equipped to coordinate well and bring uniformity in the concepts, design and development of standardised formats and implementation of data base at national level covering all gamut of subject areas in their domain of functioning.

The following three stage approach as that of NAS would help in integrating the various statistics at national level.
(i) The first, and most obvious, factor is related to the fact that one of the characteristics of 'transactions', the main units of observation in economic statistics, is the size; and that the size of transactions, as measured by a continuous variable 'value', can be added/aggregated for transactions that are defined as 'similar' according to other variables observed (e.g. the parties involved or the kind of goods and services involved, their purpose, how they are financed). Through such aggregations one arrives at measuring concepts that are meaningful for a wide range of descriptive and analytical purposes. Similarly identifying key terms which

could be associated with the subject matter and quantified in terms of smaller terms which could be aggregated at various levels is key to integration of data within the subject domain.

(ii) The second factor is that there exist reasonably clear descriptive and analytical purposes and models that have made it fairly easy to agree on criteria that would define important and interesting similarities and differences between transactions, and the aggregates that as a consequence could be measured were more interesting as units of analysis and descriptions than the individual transactions. Such agreement on subject domain should be developed.

(iii) The third factor is that the development of NAS happened through the combined efforts of users of the results and producers of the underlying relevant statistics, with the users being the main formulators of the system's specifications. This meant that there was a 'user guarantee' in the system from the start. The 'user' involvement to create a system for 'social and demographic statistics' that could have similar coordinating effects was very much needed.

The main starting points for coordination and coherence in all sphere of official statistics therefore must be: (1) that the use of such statistics is predominantly to describe and analyse micro issues, i.e. the situation and behaviour of individuals and the events and institutions that influence them; and (2) to carefully examine the way the basic elements for such statistics are constructed, i.e. the definition of population and units, reference periods, variables and classifications (value sets), as well as clear definitions of the relationships between different units of observations as basis for assigning variable values to them. Coherence in time and geographic references will be important for certain areas and issues, as will be the fact that activities and many events are using.

Having standardised the concepts and definition of unit level data, it is matter of aggregating at various levels, village, block, tehsil, district, state and nation. What is important is uniform methodology, aggregation multiplier at various levels, standard error and dissemination of data in strictly uniform format. Here the requirement of uniform and compatible Hardware, Software and well trained man power should not be ignored.

16. Availability of Official Statistics

Both Central and State Governments collect, produce, and disseminate a huge volume and variety of data in the course of their operations. While much of this data relates to government administration — to budgeting, planning, and program performance — it is official statistics that most capture our interest. These statistics —demographic, economic, and social information — collectively paint hard-data local and national portraits of people and businesses. They motivate and justify government spending and private-sector investment alike.

Official statistics cover the gamut of societal concerns. Major sectors include agriculture, commerce, education, health, housing, poverty, and transportation to name a few. Statistics are collected from censuses, surveys, and routine operational reports. They constitute a form of Public Intelligence, not only about public concerns

but also of great interest to the public. This section deals with availability of data at national level for select few but important sectors, their periodicity, their usability including to the extent possible the content, coverage, web-site and data portability.

16.1 Agricultural Statistics

In a developing country like India, where agriculture plays an important role, the need for knowledge about agricultural statistics and limitations from which they suffer has grown for the vast research community. As per the latest release, GDP from Agriculture Sector is Rs. 692499 crore registering a growth rate of 5.4 % over the previous year and contributing 14.2 % to the overall GDP (Advance Estimate for the year 2010-11 at 2004-05 price)
Agricultural statistics system in India is decentralized both horizontally and vertically. Primary statistics are collected by the State governments (provincial or sub-national) and consolidated for the country as a whole by the Directorate of Economics & Statistics (DES), under the Department of Agriculture & Cooperation, Union Ministry of Agriculture. This system, which has evolved over the course of time, provides various sets of statistics, data, indices and indicators. Agricultural statistics are also generated through various surveys and statistical operations conducted by different institutions and government departments. The DES is the nodal agency for compiling, documenting and disseminating the basic data and the key indicators at the national level.

DES releases every year estimates of production and yield of foodgrain crops, oilseed crops, sugarcane, fibre crops and important commercial and horticultural crops. Data on nine-fold land use classification, irrigation (crop-wise and source-wise) are also collected and compiled at the national and sub-national levels on an annual basis. Weekly data on wholesale/retail prices and farm harvest prices are collected from agriculture markets and used for the compilation of wholesale price index for agricultural commodities. DES produces an annual publication entitled 'Agricultural Statistics at a Glance'. The publication also covers data relating to agriculture on national income and social economic indicators, outlay and expenditure, capital formation, area, production and yield of principal crops, cost estimates, procurement by public agencies, per capita net availability, consumption and stocks, import/ export, tariff, wholesale price index, land use statistics, census of agricultural inputs, wages of agricultural workers, livestock population and fish production in country, inter-alia. This publication is based on the data collected which are compiled by the Directorate of Economic and Statistics and various Ministries and Departments in the Government of India. Agricultural Statistics at a Glance is available in the website at http://agricoop.nic.in and http://dacnet.nic.in/eands. As on date second advance estimates of principal crops for the year 2010-11 are posted in the web-site. Most of the data set available are in PDF format.

Department of Animal Husbandry, Dairying and Fisheries

This Department collects and compiles, among other things, quantitative data on livestock population and products which include cattle, poultry, wool, meat, and meat products. Further, it furnishes production data on milk, egg, wool, fish & fish seed. The website of this Department is http://dadf.gov.in. It has a web-based system for accessing State-wise and district-wise livestock census data and agricultural machinery of the country. The sources of these data are the State (sub-national) and Union Territory Governments. At the national (all-India) level, all matters relating to Fisheries are looked after by a Joint Secretary in the said Department. It may be mentioned that data on separate budgetary allocation for the Statistical Wing as such are not available. Fisheries is an important sector in India--it provides employment to millions of people and contributes to food security of the country. With a coastline of over 8,000 km, an Exclusive Economic Zone (EEZ) of over 2 million sq km, and with extensive freshwater resources, fisheries play a vital role.

Important data sources are:

- Department of Animal Husbandry, Dairying and Fisheries, (http://dahd.nic.in)
- National fisheries development board, Government of India (http://www.nfdb.org)
- Central Marine Fisheries Research Institute (http://www.cmfri.com)
- Fisheries- India portal (http://india.gov.in/sectors/agriculture/fisheries.php)

16.2 Health Statistics

The Plan Allocations to Ministry of Health and Family Welfare for the year 2011-12 is Rs.26,760 crore to improve the overall health of the Indian citizen. Huge amount of data is generated by both surveys and administrative machineries. As an offshoot of administrative procedures and implementation of various national health programmes, many data series flow into the various organisations/divisions of Directorate General of Health Services, Department of Health and these form the basis of the annual central health statistical publication *Health Information of India* brought out by CBHI. The central Programme Officers consolidate the programme-specific information flowing from States and UTs and furnish these to CBHI for publication. Data are also collected by CBHI directly from the Directorate of Health Services of all States and UTs, statutory councils such as, Medical, Dental, Nursing and Pharmacy Council, Office of Registrar General of India (RGI), and from other Central Departments and international organisations for publication. Data are presented at the State and UT level along with all-India figures.

Apart from the regular annual publication, *Health Information of India*, CBHI also brings out *ad hoc* publications. *Health Map of India, which* depicts in maps, the district wise number of hospitals with bed strength and dispensaries, as also the number of specialised hospitals in various districts. The data are also presented in tabular formats side by side with the maps. The CBHI is also the nodal agency to implement the Health Management information System (HMIS) that was started to rectify the information deficiencies in the area of implementation of various health

and family welfare programmes and by routine health service activities, as well as to provide linkages among them on a monthly basis.
The Rural Health Division of DGHS compiles and publishes *Rural Health Statistics*. This is a six-monthly bulletin, containing information on Government health infrastructure and manpower deployment in the rural areas. This publication also presents data at State and UT level. The National AIDS Control Organisation (NACO) under Department of Health collects data on cases and deaths due to AIDS/STD and publishes these in its *Annual Update*. This is one of the first Ministry to implement Data Warehousing solutions for Health Statistics.
Web-site: http://mohfw.nic.in

16.3 Education Statistics

Budgetary allocation for the HRD Ministry is Rs. 52,057 crores for the financial year 2011-12. The increased funds have been proposed to implement the government's flagship literacy programme "Right To Education Act" through the Sarva Shiksha Abhiyaan (SSA). The Educational Statistics that are collected annually from educational institutions and are published at the national level by the Ministry of HRD. The statistics collected annually from schools for the primary level of education in the districts covered under the District Primary Education Programme (DPEP). The DPEP is a Centrally sponsored scheme providing a special thrust to achieve universalisation of primary education through decentralised management, participatory processes, empowerment and capacity building at all levels.
The DPEP-EMIS is an important source of data on enrolment, teachers, physical facilities, etc. for the primary level of education, but since it is confined to the DPEP districts and the statistics at present remain unpublished, it cannot be regarded a part of the country's statistical system as yet. However, the data is compiled annually and is made available generally within a year. In addition to the above, there are other organisations that provide data on Education Statistics. These are given below:
Directorate of Employment & Training for data on the educational level of the job seekers through Employment Exchanges;
National Family & Health Welfare Survey, for data on literacy and children attending school, based on a sample survey of households;
Database Report on Vocationalisation of School Education survey;
Database created by the State Governments; and
IAMR Project on National Technical Manpower Information System;
With regard to medical and dental education, the data available are meant for the specific purpose of regulating admissions to the MBBS and PG courses for Allopathic and Dental systems of medicine. Department of Agricultural Research and Education (DARE) is collecting data through various schedules.
The NCTE has a computerised Management Information System (MIS) to maintain limited data on teacher education collected through the Performance Appraisal Report.
Web-sites: www.education.nic.in

16.4 Labour Statistics

The Labour Statistics in India are largely collected under various labour laws and regulations through the administrative system, even though a large portion of the workforce in India is engaged in agriculture and informal sector. Collection of timely, reliable and adequate data on labour sector and its timely dissemination to users requires immediate attention, to bring desirable improvement in the system and to meet the data requirements of the planners

The DGE&T brings out a number of publications based on the data collected through the Employment Market Information Programme (EMIP) and the National Employment Service (NES). The data collected through EMIP is disseminated through various publications, which provide estimates of the utilisation of labour force in different sectors, industries, occupations, etc. the excess and shortage of manpower and the present level of employment generation in various industries. The publications of DGE&T include *Quarterly Employment Review*, *Quick Estimates of Employment in the Organised Sector* (Quarterly), *Employment Review* (Annual), *Occupational-Educational Pattern of Employees in India* (for public sector and private sector in alternate years), *Employment Exchange Statistics* (Annual), *Census of Central Government Employees* (Annual) and *Bulletin of Job Opportunities in India* (Annual).

Web-site: http://labour.nic.in

16.5 Industrial Statistics

Manufacturing sector recorded a GDP of Rs. 776337 crore with 8.8 percent Growth Rate at 2004-05 price during the financial year 2010-11 as per the latest Advance Estimate of GDP. ASI is the major formal source for a comprehensive database on the different aspects of industrial activity in the country. For the purpose of collection of data relating to manufacturing activities through sample survey, all manufacturing units in the country are classified into two broad sectors namely, registered and unregistered sectors or organised and unorganised sectors–the terms being quite often used interchangeably. While the registered manufacturing sector covers the manufacturing units registered under sections 2m (i) and 2m (ii) of the Factories Act, 1948 or under the *Bidi* & Cigar Workers (Condition of Employment) Act, 1966, i.e. the units employing 10 or more workers and using power or 20 or more workers but not using power, the unregistered manufacturing sector covers all residual manufacturing units.

Data with respect to the unregistered manufacturing units or enterprises are collected through the periodic sample surveys. Apart from these sample surveys specifically designed to collect detailed data relating to employment, fixed assets, working capital, details of input and output, gross value added, etc. from the unregistered manufacturing enterprises, information on number of workers engaged in manufacturing activities are also available from three more sources namely, Employment-Unemployment Surveys (EUS) of the NSSO, decennial Population Censuses and the periodic Economic Censuses. The Economic Censuses also give information on the number of enterprises. It may be mentioned that employment data available from each of these three sources takes into account all types of workers irrespective of whether the employing enterprises are registered or

unregistered in nature. Also MOSPI releases the Index of Industrial Production(IIP) every/month to indicate the growth of organized manufacturing sector.
Web-site: http://mospi.nic.in/

16.6 Trade Statistics

Exports of goods and services during the year 2010-11at constatnt price (2004-05) is Rs. 1101148 crore with 20.6 percent growth rate as per the latest Advance Estimates and Imports is at Rs. 1419365 crore with 26.6 percent GR. Trade Statistics are obtained as a by-product of administrative activity. In the case of External Trade, there are three stages of administrative activity namely, licensing, actual shipment and arrival of goods, and the receipt and remittance of payments. The Director General of Foreign Trade (DGFT) is responsible for licensing statistics; the Director General of Commercial Intelligence and Statistics (DGCI&S) for the balance of trade statistics and the Reserve Bank of India (RBI) for the balance of payment statistics.
DGFT Web-site: http://dgft.delhi.nic.in
RBI Web-site: http://www.rbi.org.in

16.7 Banking statistics collected

Business of all scheduled banks in India: Scheduled commercial and cooperative banks statutorily submit to RBI a fortnightly return on major items of its assets and liabilities in India as also a special return for the last Friday of every month if it is not a reporting Friday (provisional data within one week followed by final data within 20 days). In order to meet the data needs in the wake of liberalisation and deregulation in the financial sector, as recommended in the Reddy Committee Report (1998), scheduled commercial banks are also required to submit more granular data on their assets and liabilities along with the statutory return which includes paid-up capital, reserves, details of foreign currency liabilities and assets, details of investments in non-SLR securities, etc. RBI publishes the provisional aggregate data on assets and liabilities of the banking system through weekly press communiqué and in the Weekly Statistical Supplement to RBI Bulletin (WSS). The final data are published in the RBI Bulletin, every month. Such information are useful for monitoring the compliance of statutory cash reserve ratio (CRR) and statutory liquidity ratio (SLR) by the banks and also for compilation of money supply. In addition, these data are also used for monitoring overall banking business in India, including the trends of important indicators like aggregate deposits, bank credit, investments, etc. The banks also statutorily submit to RBI their annual reports (balance sheet and profit and loss account). These data are published in the various publications of RBI. Detailed bank-wise data based on annual reports are published in 'Statistical Tables Relating to Banks in India'.

Basic Statistical Returns (BSR): These are a set of seven non-statutory returns in place since 1972 aim to capture a detailed picture of the distribution of bank credit, deposits, investments, etc., across various dimensions. **BSR-1** collects branch level detailed data from SCBs and RRBs on gross bank credit including term loans, cash credit, overdrafts, bills purchased and discounted, bills re-discounted, etc. from scheduled commercial banks and regional rural banks on annual basis (reference

date: 31st March). BSR-1 return is divided into two parts - Part A and Part B (termed as BSR-1A and BSR-1B). In BSR-1A, for credit limit above Rs.2 lakh, account-wise information is collected on various characteristics, such as centre and population group of utilisation of credit, type of account, type of organisation, occupational category, nature of borrower account, rate of interest, credit limit and amount outstanding. In BSR-1B, information in respect of small borrower accounts (credit limit up to Rs.2 lakh) is obtained in consolidated form for broad occupational categories. **In BSR-2**, branch level information from SCBs on deposit accounts (excluding inter-bank deposits) with their break-up into current, savings and term deposits and various other characteristics such as gender, maturity period, etc., are collected. BSR-2 also contains information on staff strength, classified according to gender and category (i.e. officers, clerical and subordinates). BSR data at various levels of aggregation are disseminated through annual publication of RBI "Basic Statistical Returns of Scheduled Commercial Banks in India". **BSR-4** collects annual (as on March 31) data on composition and ownership pattern of bank deposits from SCBs through a sample survey of branches. An article presenting salient features of the results of the BSR-4 survey is published annually in the RBI Bulletin. **BSR-5** collects annual (as on March 31) data on investments of banks in India and abroad to capture changes in composition pattern of investments of SCBs according to type, maturity profile, interest/coupon rates, etc. An article presenting salient features of the results of the BSR-5 survey is published annually in the RBI Bulletin. **BSR-6** collects data on 'Debits to Deposit Accounts' for the SCBs on a quin-quennnial basis based on a sample survey of bank branches. BSR-6 data are used to estimate the rate of turnover of deposits, which is one of the important measures of economic activity in the country as a whole. An article presenting salient features of the results of the BSR-6 survey is published in the RBI Bulletin. **BSR-7** collects branch/office-wise information on aggregate deposits and gross bank credit of SCBs (including RRBs) on quarterly basis (last Friday of June, September and December quarters and March 31) from the head/controlling offices of banks. The data are disseminated through 'Quarterly Statistics on Deposits and Credit of SCBs in India".

International Banking Statistics - At present, banks in India provide various details of their operations in India as well as abroad to different departments in the RBI to meet the specific requirement of the departments concerned. The IBS system of the Bank for International Settlements (BIS) is designed to collect/compile/provide information on banks' external/international liabilities and assets vis-à-vis (a) non-residents in any currency and (b) residents in foreign currency. Under the system, information from SCBs on loans, deposits, investments, borrowings, other assets and other liabilities with details into currency (domestic and foreign currencies), sector (banks, non-bank public and non-bank private) and country (individual countries, international institutions and monetary authorities) is compiled at quarterly intervals. Since March 2001 quarter, the consolidated data of banks in India in the form of 23 statements (18 locational banking statistics (LBS) and 4 (6 from December 2010) consolidated banking statistics (CBS)) are being supplied regularly to the BIS on a quarterly basis. BIS started incorporating India's IBS data from December 2001 quarter and as a result, India became the third among all developing countries in the world complying with the BIS requirements of compilation of IBS. While BIS publishes consolidated data of all reporting countries,

RBI publishes consolidated data on IBS of India in the form of article in the RBI Bulletin.

Branch Banking Statistics: RBI collects data on different aspects of banks through periodical returns/ statements. For processing these data, it is necessary to keep a unique identity of the source of data. This is achieved through allotting suitable code number, named as Uniform Code Numbers to all the bank offices. Comprehensive and updated list of branches is maintained by RBI in the Master Office File (MOF) constituting the frame of bank branches for various BSR surveys, other bank related surveys and various foreign exchange related returns received in RBI. Basic data sources are commercial banks, cooperative banks and non-banking financial companies handling foreign exchange business, wherefrom branch banking details are obtained.

Data on Interest Rates - Data on interest rates offered on deposits/charged on credit by the scheduled commercial banks are collected by RBI on a regular basis. More detailed data on range of interest rates in respect of different types deposits and credit are collected on fortnightly/monthly/quarterly basis from the SCBs in pre-defined formats. Statistics on flow of gross bank credit to different sectors of the economy, viz.; agriculture, industry, etc., is collected and compiled through BSR-1 return on an annual basis.

Sectoral Deployment of Credit - In addition to the annual data collected through BSR system, RBI collects provisional information on sectoral deployment of credit from select banks on a monthly basis (as on last reporting Friday) in view of the need for such information on more frequent intervals and with minimum time lag. The consolidated information based on the data is published in various publications of the RBI.

Priority Sector Advances - Besides statistics based on statutory and special returns, RBI collects and compiles various other banking statistics relating to priority sectors. Information based on these subjects are collected half yearly through various returns and are regularly disseminated in various publications like Statistical Tables Relating to Banks in India, Report on Trend and Progress of Banking, etc. In order to align bank credit to the changing needs of the society, the scope and definition of priority sector have evolved over time by including new items as also by enhancing credit limit of the constituent sub-sectors. The coverage of the priority sectors, the data, which are published in its various publications by RBI are agriculture, small scale industries, small road and water transport operators, professional and self employed persons, education etc.

Money Supply Statistics – RBI has a long tradition of compilation and dissemination of monetary statistics, since July 1935. Monetary statistics at present are compiled on a balance sheet framework with data drawn from the banking sector, viz., scheduled commercial banks, cooperative banks, urban cooperative banks and postal authorities. The rationale and analytical foundations behind the compilation of monetary aggregates have been provided to the public through various reports, especially through the reports of the various working groups. Monetary aggregates are compiled on fortnightly basis and published regularly in

most of the major publications of RBI, such as the Annual Report, Report on Currency and Finance, Handbook of Statistics on Indian Economy, RBI Bulletin, Weekly Statistical Supplement, etc.

A list of regular RBI Publications include:

Annual

- Annual Report
- Report on Currency and Finance
- Report on Trend and Progress of Banking in India
- Handbook of Statistics on the Indian Economy
- State Finances: A Study of Budgets
- Statistical Tables Relating to Banks in India
- Basic Statistical Returns of Scheduled Commercial Banks in India

Quarterly

- Macroeconomics and Monetary Development
- Occasional Papers
- Quarterly Statistics on Deposits and Credit of Scheduled Commercial Banks

Monthly

- RBI Bulletin
- Monetary and Credit Information Review

Weekly

- Weekly Statistical Supplement

A Central Resource: the RBI's Database for Indian Ecopnomy

- Enterprise-wide data warehouse
- User-friendly, public access via RBI web site, www.dbie.rbi.org.in
- Pre-formatted reports
- Simple and advanced queries
- Definitions of basic concepts

Web-site: http://www.rbi.org.in

16.8 Socio-Economic Statistics

Socio-economic Statistics thus form an important component in the development of the country and include a vast array of information on health and disease; literacy and education; standard of living and poverty; labour force and employment; status of women and gender empowerment; population parameters relevant to fertility, mortality and migration; ecology and environmental protection. Timely collection of appropriate, adequate and reliable data on the above dimensions and proper use of this in planning, implementation, monitoring, evaluation and redesign of various developmental programmes is absolutely essential if the country has to progress more rapidly and join the ranks of the developed countries in the near future. Considerable amount of data is generated in line Ministries and MOSPI. Some of them like Labour, Health, Education, MOSPI etc are already covered. Ministry of Social Justice and Empowerment, Ministry of Women and Child Development,

Planning Commission, Registrar General of India, Ministry of Tribal Affairs, Ministry of Minority affairs, etc. generate voluminous data on various socio-economic spectrum of the people.
Related web-sites are:
MOSPI: http://mospi.nic.in
M/o Social Justice & Empowerment: http://socialjustice.nic.in
M/o Women and Child Development: http://www.wcd.nic.in
Registrar General of India: http://www.censusindia.net
Ministry of Tribal Affairs: http://tribal.gov.in
Ministry of Minority Affairs: http://www.minorityaffairs.gov.in

16.9 Revenue Statistics

The budget presents only the aggregate heads of major taxes both at the level of the Centre and State. The estimates for State-wise devolution of taxes and duties are disseminated through the budgets. Details regarding tax refunds, tax arrears, accruals, cumulative collected for the previous years are published in the web-site hosted by CBEC and CBDT. The *All India Income Tax Statistics* (AIITS), an annual publication of the Income tax Department is based on a very small sample size and published with a considerable time lag. Tax collected by union government on Customs, Central Excises, Service tax details, etc. are available on monthly basis in
CAG web-site: http://www.cag.gov.in
Income Tax: http://www.incometaxindia.gov.in
Central Excise and Customs: http://www.cbec.gov.in

16.10 ENVIRONMENT STATISTICS

The Ministry of Environment and Forests is maintaining information on various aspects related to environment. As environment is a multi-disciplinary subject involving complex subjects like bio-diversity, atmosphere, water, land and soil and human settlement, it is very difficult to collect and analyse data and study inter-relationships. As there are several agencies for collection of information, it becomes necessary to develop an efficient statistical system on environment that could meet the growing demand of various stakeholders, both Government and non-Government as well as outside agencies for environmental data.
Web-site: http://moef.nic.in

16.11 Price Statistics

The data on prices are regularly collected by Central and State Government Departments and agencies for varied purposes. These data basically form the source of varied information compiled in different forms in accordance with the specific needs of multiple agencies.

Web-sites:

For CPI(IW) & CPI(Rl/AL) – Labour Bureau: http://labourbureau.nic.in
For WPI – OEA : http://eaindustry.nic.in
For new series of CPI – MOSPI : http://mospi.nic.in

16.12 corporate sector statistics

The responsibility for collection, compilation, maintenance and dissemination of basic statistics on the Indian Corporate Sector is vested with the Ministry of Company Affairs (MCA). The registered companies are required to file certain documents and returns with the Offices of various Registrars of Companies (ROCs) under the provisions of the Companies Act. The most important of these are the Annual Reports and Balance Sheets of the companies and returns on share capital. Thus, the Corporate Sector Statistics maintained by the MCA are basically a by-product of the administration of the Companies Act. A consolidated list of all the newly-registered companies in the year with their names, addresses, industrial activities and authorised capital is available month-wise. The distribution of companies by various categories such as Government and non-Government, public and private, State of registration and industrial activity is available. The capital raised by the existing companies is available on a quarterly basis.
Web-site: http://www.mca.gov.in

16.13 National Accounts Statistics

The overall GDP for the financial year at constant price (2004-05 price) is Rs. 4879232 crore with 8.6 percent GR for the financial year 20010-11. The Central Statistical Office (CSO) in the Ministry of Statistics and Programme Implementation (MoSP&I) is responsible for the compilation of NAS. At the State level, State Directorates of Economics and Statistics (DESs) have the responsibility of compiling there State Domestic Product and other aggregates.
The aggregates compiled and released (at current and constant prices) at annual periodicity by the CSO include gross and net domestic product by economic activity, consumption, saving, capital formation and capital stock, public sector transactions and dis-aggregated statements, as well as the four consolidated accounts of the nation namely, (a) Gross Domestic Product and expenditure, (b) National Disposable Income and its appropriation, (c) Capital Finance; and (d) External transactions. The CSO also releases the quarterly GDP estimates on the last working day of each quarter, the estimates pertaining to the previous quarter. In addition to these macro-aggregates, the CSO compiles and releases the Input Output Transaction Table (IOTT) at a periodicity of 5 years and the State-wise and crop-wise value of the output of agricultural crops at an *ad hoc* periodicity. The CSO also maintains the database on Gross and Net State Domestic Products, by industry, which is compiled by the State DESs.
Web-site of MOSPI: http://mospi.nic.in

Data Dissemination Policy

(Annexure-3)

No: P. 12011/4/92-NSSO
Government of India
Ministry of Planning & Programme Implementation,
Department of Statistics
****

Sardar Paltel Bhawan, Parliament Street
New Delhi, dated the 6th January, 1999

O F F I C E M E M O R A N D U M

Subject: National Policy on dissemination of statistical data.

The Cabinet, in its meeting held on 9th September, 1998, approved the proposal the Department of Statistics for a **"National Policy on Dissemination of Statistical Data"**. The details of the policy are given below:

(1) Dissemination of official statistics in the form of reports, ad-hoc and regular publications, etc., as at present should continue. Validated data, though unpublished, including unit/household/establishment level data after deleting their identification particulars to maintain confidentiality should also be made available to the national and international data users in the form of hard copies and on magnetic media on payment basis;

(2) No data, which are considered by the concerned official data source agency to be of sensitive nature and the supply of which may be prejudicial to the interest, integrity and security of the nation, should be supplied. The Central Government, or a State Government or the concerned Government agency, as the case may be, shall exercise its overriding prerogative to decide the degree of sensitivity of the official statistics produced by it. The data source agency will reserve the right on whether to withhold its release altogether or to release selectively;

(3) Price of data to be supplied under (1) above should include the cost of stationery, computer consumables and computer time for sorting information. However, cost of collection and validation of data will not be charged. Postal charges to be included along-with cost of data supplied.

(4) Price may be fixed in Indian Currency (i.e. Rupees) as well as in Sterling Pound and American Dollar. Foreign currency prices may be determined using relevant official multiplier fixed from time to time for printed Government publications;

(5) Survey results/data should be made available to the data users in India and abroad simultaneously after the expiry of 3 (three) years from the completion of the field work or after the reports based on survey data are released, whichever is earlier;

(6) Data users will give an undertaking in the prescribed form to the effect, inter-alia, that the official statistics obtained by him for his own declared use will not be passed on with or without profit to any other data user or disseminator of data with or without commercial purpose;

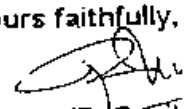
(7) Data users will have to acknowledge the data sources in their research work based on official statistics. One copy of the research study along-with short summary of conclusions, if required by the concerned data source agency, should be supplied in the form of hard copy or on electronic media, free of cost; and

(8) The Department of Statistics will be the nodal agency for dissemination of official statistics produced by Central Government Ministries and Departments. However, the concerned subject matter Ministries and Departments of the Central Government will be the final authority on issues arising out of this policy with a view to resolving any dispute between a data user and a data source agency.

2. It has also been decided that the exact delineation of the role of the nodal agency will be decided by Committee of Secretaries. The principles of determining the pricing of data may be varied from time to time by the Department of Statistics, in consultation with the Ministry of Finance.

3. A separate communication will follow with regard to the modalities of implementation of the above policy.

Yours faithfully,



(R. Ravi)
Under Secretary to the Government of India
Tel: 374 7503
Fax: 334 2384

File No. P-12011/4/92-NSSO
Government of India
Ministry of Planning & Programme Implementation
Department of Statistics & Programme Implementation
****

Sardar Patel Bhavan, Parliament Street,
New Delhi, dated the 3rd May, 1999

OFFICE MEMORANDUM

Sub : National Policy on Dissemination of Statistical Data.

The Office Memorandum of even number dated 6th January 1999 on the above subject may be referred to. As mentioned therein, the Cabinet while approving the policy, had directed that "Exact delineation of the role of the nodal agency will be decided by a Committee of Secretaries. The principles of determining the pricing of data may be varied from time to time by the Department of Statistics in consultation with the Ministry of Finance". In pursuance of the above, the following guidelines have been approved by the Committee of Secretaries:-

i) Dissemination of official statistics in the form of reports, ad-hoc and regular publications, etc. by the Central Ministries/Deptts/Agencies as at present shall continue

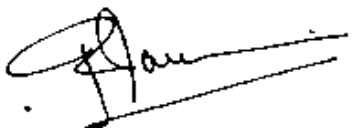
ii) A data warehouse in the Department of Statistics will be created to enable the data users and general public to have easy access to the published as well as unpublished validated data from one source.

iii) The data warehouse will collect data from various source agencies, integrate the data into logical subject areas, store the data in a manner that is accessible and understandable to non-technical decision makers and deliver data/information to decision makers through report writing and query tools.

iv) As data source agencies are generating data at various levels, the responsibilities of data supply and receipt will be shared between the respective Central Ministries/Departments/Agencies and the Department of Statistics by establishing and maintaining close collaboration.

v) For each data type and source, detailed studies will be undertaken by the Department of Statistics in co-operation with the concerned data source agency on (a) the concepts, definitions, classifications and methods used in data collection and processing including validation, (b) formats of data generation, (c) media on which data will be supplied, (d) frequency of supply of data and (e) procedures and modalities for preservation, updation and dissemination of data.

vi) The volume of data flowing from each source agency into the data warehouse will be assessed by the Department of Statistics in order to formulate the various parameters required for designing, establishing and maintaining a data warehouse.



vii) Each data source agency will be required to adopt for itself a calendar for preparation and release of data which it will share with the Department of Statistics. As a part of its nodal responsibility of dissemination of data from the source, the Deptt. of Statistics will keep track of the data release calendar of each source agency.

viii) The data source agency will be required to supply on computer compatible media validated data, published or unpublished free of cost to the data warehouse.

ix) The Department of Statistics will prepare Directories of all available data in the data warehouse and update the same at frequent intervals. A web-site will be created for the data warehouse and the Directories will be available on the web-site.

x) From the data warehouse, data/information will be made available free of cost to the data source agencies for official use and also the approved research institutes and universities for research purposes.

2. The price of data to be supplied to users other than those mentioned in para 1(x) above will depend upon system of hardware and software used for data storage, retrieval, sorting of information etc., and also on medium of supply of data. The cost will be determined on setting up of Data warehouse in consultation with Ministry of Finance by the Department of Statistics.

3. In view of para 1(i) above, the pricing policy in respect of the existing published data followed by the Departments will be continued/decided by the Department on their own.

(R. Ravi)
Under Secretary to the Government of India
(Tel: 374 7503)
(Fax: 334 2384)

To

(1) Secretaries to all Departments/Ministries in the Government of India

(2) Chief Secretaries to all States/Union Territories