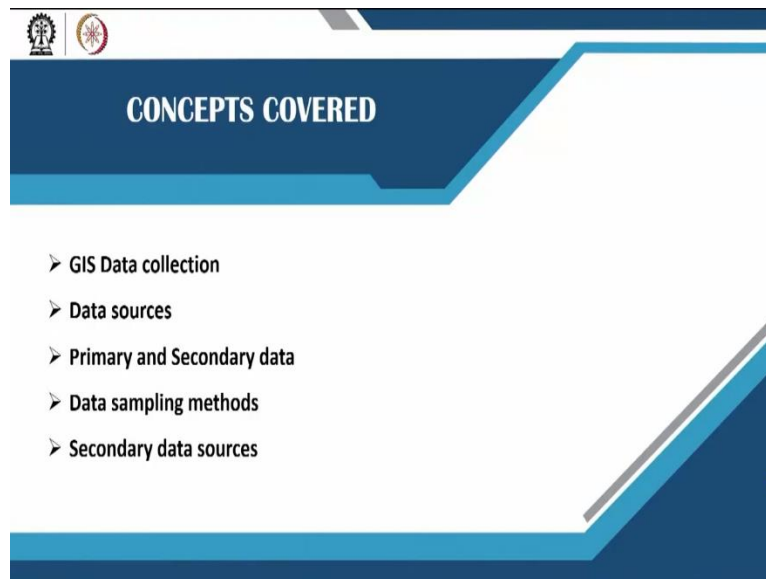


**Geographic Information Systems**  
**Prof. Bharath H Aithal**  
**Ranbir and Chitra Gupta School of Infrastructure Design and Management**  
**Indian Institute of Technology – Kharagpur**

**Module No # 02**  
**Lecture No # 09**  
**GIS Data (Continued)**

Hello Namaste, we welcome back to the course we are in the lecture 4 of the module 2. I did explained in the last class that we will be probably working on GIS Data modules as such or looking at GIS data models, but I would first way I get into how do you connect data then get into data models. So, I would go stepwise so first is how do you would collect the data? What kinds of methods you have to follow? Where and how you get the data? So I have given you lot of sources already that where the data is available. But I will inform you how the data is collected.

**(Refer Slide Time 01:03)**



So when I look at the entire concepts that I would be speaking on in this class the first thing is look at the GIS data collections. What are different data sources that are there? It may be primary data or the secondary data sources. And how the primary data is efficient and the secondary data sources there are certain issues that has to be accounted before you consist the secondary data sources to build the data models then data sampling methods.

So this is where most of us make a mistake. We do not consider at definitive sampling methods. We have we do only understand that if this is how it goes then the sampling can be done. But we should look at how the data is collected in a proper way. So the data sampling methods is extremely important and the way it is collected the way you understand your process then that makes a different model. Then finally the secondary data sources how to handle the secondary data sources also we look at it and this slides.

**(Refer Slide Time 02:08)**

The slide is titled "GIS Data Collection" in blue text at the top left. The background features a stylized tree with various icons (gears, a hard hat, a beaker, a laptop, a globe, etc.) as branches. The text on the slide is as follows:

- Obtaining data is an important part of any GIS project
- You need to know
  - What types of data you can use with GIS
  - How to evaluate it
  - Where to find it
  - And how to create it yourself

At the bottom left, there are logos for IIT Bombay and NPTEL. At the bottom right, the text "IT Change" is visible.

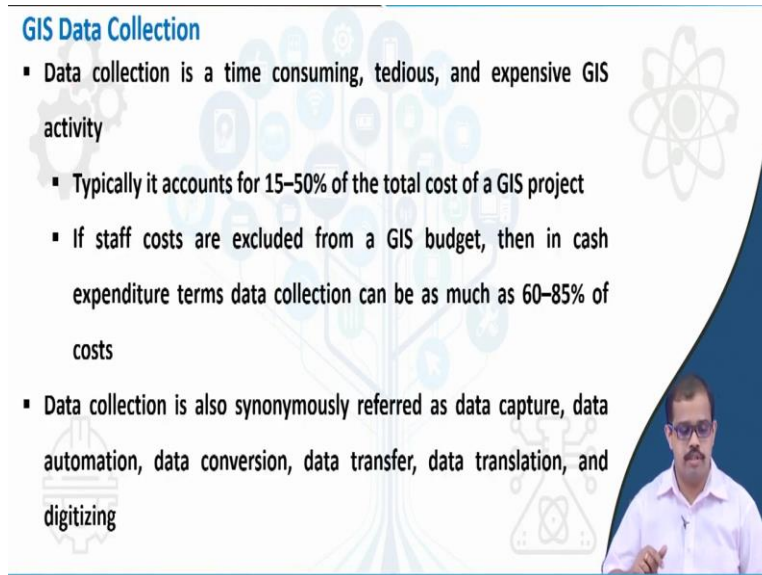
So when we look at GIS data collection, so when the most important part of any GIS project is data, the collection of data. So if you ask any GIS user or any user who is working on GIS data they would said it is very difficult to get a data. But why it become very extremely difficult to get a data is probably we do not know the source of data or we do not know how we actually generate the data. How do we handle a data?

So what we know need to understand here is what types of data you can use with GIS. It can be any type of data, but how to evaluated? How do you actually use it? How do evaluated for being a better data? Where do you find the data? So I have given huge number of examples, please look at all those website which I was talking about and my previous module. So look at the website that gives you extreme understanding about how data is there.

India has huge collection of data Indian subcontinent can give you huge data sets. So please look at the data and different source of data. Then how do you create your data its yourself? You can

create your data yourself for your project. You do not need to depend on any other sources. So you have to create it. So how do you create it? Which means how do you sample it first of all. So that is what very important, that is what we would understand in this particular lecture.

**(Refer Slide Time 03:43)**



**GIS Data Collection**

- Data collection is a time consuming, tedious, and expensive GIS activity
  - Typically it accounts for 15–50% of the total cost of a GIS project
  - If staff costs are excluded from a GIS budget, then in cash expenditure terms data collection can be as much as 60–85% of costs
- Data collection is also synonymously referred as data capture, data automation, data conversion, data transfer, data translation, and digitizing

The slide features a background with faint icons of a globe, a smartphone, and a network diagram. A small inset video in the bottom right corner shows a man with glasses and a mustache, wearing a light blue shirt, speaking.

So when we always look at as I said data collection itself is time consuming. And many of the projects if you see data collection itself is 15 to 50% of the cost in any GIS project. So if you think about the cost if you just take out the staff cost the person who is actually collecting the data almost 80% of the entire the cost is in terms of data collections. Almost 80 to 85% so anyone who is working with GIS will definitely tell you that the data itself is what matters.

But you can get data yourself it may not even cost more because it will cost more only you do not know the sources of actually getting the data. So when we look at data, data collection is also synonymously referred to as data capture in many ways data automation, data conversion, data transferred, data translation, and data digitalization. These are the different ways of saying how the data is actually used or data is actually acquired, or the data is collected ok.

**(Refer Slide Time 04:58)**

## Data Sources

### ▪ Two types of data sources

#### 1. Primary data

- Data measured directly by surveys, field data collection, remote sensing

#### 2. Secondary data

- Data obtained from existing maps, tables or other data sources

Data Source	Raster	Vector
Primary	Digital remote sensing images/satellite data	GPS data collection
	Digital aerial photographs/Photographs from UAV	Surveying data collection
Secondary	Scanned maps/topographic maps	Topographic surveys/Lidar survey
	Digital elevation model from maps	Taxonomy and other attribute data from existing datasets



So when I speak about data there are two types of data one is the primary data other one is the secondary data. So when I say primary data these are the data that is measured by you through surveys, field data collections or may be some of the remote sensing data that you may get in. With advent of UAV's your service and field data collection has become extremely easy. And only thing is that probably the government has mandated the use of licenses in terms of UAV's and only in the specific regions.

So that primary data collections has become extremely easy, but if you have primary data then your data model representation would be extremely good. So whenever you are building up your data model try to collect as much as primary data as possible that is my advice. So visit the place so there are many examples of various students doing the analysis without even visiting the site or visiting the place of which they are trying to and or making into the data model.

If you do not understand the properties of site if you do not visit the site or if you do not see the site at all how it is extremely difficult to mimic the same site in the data model. So the first thing is look at the data look at the primary surveys, go to the data, go get it get the primary data. So that is extremely important. Next is the secondary data, so data obtained from the existing maps ok, tables or any of the other government there is huge number of government sources which gives you data. It may be statistically, it may be quantitatively, it may be qualitatively.

So what are the different sources, all of this sources becomes secondary source ok. So for example I have mentioned some of the data sources here. If you look at the raster data model in the primary source, you have digital remote sensing data or it is also called a may be a satellite data. Or the any remotely sensed data, aerial photograph, photographs from UAV's. So when you look at a primary data as a vector data model you have GPS data collection and surveying data collection that can be in form of vectors.

So secondary data when you look at it scanned maps and topographic maps i will give you some example of this. Digital elevation models with maps that also can be your secondary data for a using a raster data model. Whereas when you are looking at the vector data model it is more of topographic survey or lidar survey and taxonomy and other attribute data from the datasets, from the governmental dataset or maybe from many of the datasets.

If you are trying to get gather more information, there are huge number of literature which can give you certain information on certain data. So that is the secondary datasets that you collect. But always secondary datasets has to be evaluated so that you are you know that information whatever is presented here that is correct. How do we know it? Using a metadata I will come to that bit later.

**(Refer Slide Time 08:09)**



So how do we do a data collection? The way the data has to be collected has to be something like this. It is a five-path process what I call it as, you have to look at first you have to plan it. If you

do not plan how to collect a data, then you are going to lose on every front of data acquisition ok then the preparation of a data. So for example when I say planning you should know what every user requires. If you are targeting a certain user set of people who will try to use that tool in order to develop some science or use that tool to deliver some aspects of application, you should know what are the user requirement.

So based on those user requirements you create you will actually populate that data ok. Then available resources that is extremely important you cannot go out of bounds to collect the data. Then you have preparation is the next step where you obtain preliminary data then look at if you have collected the data then collect then you have to look at the data cleaning. Which means you are correcting actually you are correcting the data for any of the non-datasets that has to be there.

Or you are looking at the noise removal and setting up the entire GIS environment. If you can do certain amount of digitizing and data transfer that is the next step of how you actually create a data through surveys. So then you have editing and improvement in case there are certain issues I spoke about the outsource I will also show you some of the examples of the how errors can be there. So editing and improvement of those specific data has to be done and most importantly is validation of the datasets.

If there are certain data sets which are secondary nature, please validate that dataset. Without validation never use the dataset which can create the issues to you when you are actually learning the model. And evolutionary and final formatting this is where identifying a process success or an accuracy assessment become extremely important.

**(Refer Slide Time 10:25)**

## Primary Data

- Resolution is the key consideration while collecting raster data (Spatial, spectral and temporal)
- We cannot usually observe the spatial distribution of a variable throughout the study area
- Therefore we need to sample:
  - Take measurements of a subset of the features in the area that best captures the actual spatial variation



But when you look at primary data as I said in my previous lecture resolution becomes a key consideration when you are collecting any of the primary data especially in terms of a raster. If you are looking at a raster that is a pixel model, then you are looking at the resolution specifically. When I say resolution it probably when you look at the raster data model, I will speak about what do you mean by spatial, spectral, and temporal.

So these are very important in collecting a primary data. So we cannot usually observe the spatial distribution of a variable throughout the study area. So it is you know you have to collect the sample data. If your study area is entire KMA you cannot go every pixel by pixel every pixel if you are lands at 30 meter by 30 meter you cannot run around with 30 meter pixel because it is absolutely impossible to collect the entire dataset even maybe even with 100, 200 or 300 people around collecting the dataset.

So you have to start sampling the way you sample that is what is the extremely important therefore we when we also collect, they should also capture the actual variation that is where the sampling becomes very important. So you cannot if you have data which is 30 meter by 30 meter and you say that I will sample at maybe 1000 meter or 2000 meters ones and or let say it is 10 kilometer by 10 kilometer one sample I do and another sample at 10 kilometer may be two three samples then I will finish the survey.

You may not be able to capture the entire variation of the entire dataset. So you have to look at what is the size? What is the resolution of the data? What is the size of the data? How you are data has been translated into the data model then look at what may be the spatial variation that you have to actually capture. So a look all of this look at all of this property before you look at how you get into the primary data collection.

**(Refer Slide Time 12:35)**

**Sampling**

- The sampling density determines the resolution of the data
- Samples taken at 1 km intervals will miss variation smaller than 1 km
- Standard approaches to sampling:
  - Random
  - Systematic
  - Stratified

	Random	Stratified	Systematic
Point			
Line			
Quadrant			

The very important facts that as I am speaking is this sampling. So when we look at sampling it determines the density the sampling density determines the resolution of the data that you collect ok. So when we look at the sample that is taken at 1 kilometer intervals it may not be able to capture anything that changes below 1 kilometer. So if you have something that you have captured in 100 meters or 200 meters it will be able to give you some changes the variations in between 100 and 200 meters.

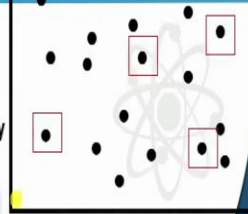
But it cannot capture less than 100 meters or 200 meters the way you are trying to capture the data. So look at what kind of a capturing of or sampling that need. What is the smallest size that you have to capture and what is the largest size that you have to capture? What are the different variations in your dataset? If so if you are trying to that then you have three types of different sampling one is random sampling, and the one is system and the other one is stratified. So we will look at the each of these sampling in detail.

**(Refer Slide Time 13:43)**



## Random Samples

- Every location is equally likely to be chosen
- Applicable when population is small, homogeneous & readily available
- Each element of the frame thus has an equal probability of selection
- A table of random number or lottery system is used to determine which units are to be selected
- Disadvantage: Not applicable for large dataset, minority subgroups of interest might be ignored and not present in population



For example the first thing is random sample so every location is equally likely to be chosen. So you can go into any location in the entire Kolkata region calculate take a sample and populate it. So this kind of sampling can be done when you have when your dataset is small ok. And it is extremely homogenous. So if you are sampling for different the way of growth mixed land use and urban land use is there so you should be able to capture the entire the different data points in the different datasets properly.

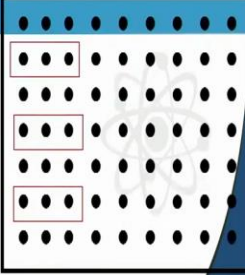

For example, if your entire area is just urban area and very minimal change then may be the random sampling is the best way to sample it. But if you have mixed land uses all across your region the way it is not homogenous is completely heterogeneous then your sampling cannot be a random sampling ok. So normally how people decide which region to be sampled or which point to be sampled is by generating a random number or just like a lottery system.

So the most used method is easier by generating a random number in the set of selected points. So the biggest disadvantage as I explained it is not capable for any large dataset. If you trying to apply for a large dataset where the way you have sampled may really go wrong or it may not give you certain subgroup that you may be extremely interested in or some of the users using that particular tool may be interested in so that is very important.

**(Refer Slide Time 15:34)**

### Systematic samples

- Sample points are spaced at regular intervals
- Involves a random start and then proceeds with the selection of every  $k^{\text{th}}$  element from then onwards
- $k = (\text{population size}/\text{sample size})$
- Starting point is not automatically the first in the list, but is instead randomly chosen from within list
- Example: Select every 10<sup>th</sup> registration number and corresponding marks of students who appeared for GATE exam
- Disadvantage: Strategy may be biased, difficult to assess precision

Then the other way is systematic sampling. So when you look at this every sample point are at regular intervals. So this involves a random start and then proceeds with the selection of every  $k^{\text{th}}$  element. So every  $k^{\text{th}}$  element is from that random sample you decide. It may be third element fifth element tenth element. Then how do you calculate  $k$ ? It is entire population size which means entire size of how you want the sample ok. That is entire the population size that you are considering now divided by the sample size.

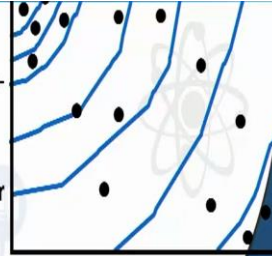
So that is the extremely important number of points said you have sampling ok divided by your sample size. So if you are trying to look at this the main disadvantage that you may come across is it may be very biased. You may be looking at only a specific aspect, but you are biasing your entire dataset to understanding only a specific aspect. For the others may not be understood in a clear way. So for example to give you a very fair example it is every tenth registration number and a corresponding student corresponding marks of the student who appeared for the gate exam.

So it does not mean that every 10 student has scored extremely well or every 10 student has scored extremely bad. So it is just a kind of sampling so it may be biased. So it may be your intuition that every 10 student extremely good student. So that biased test may should not be there in your dataset. But how do you actually move this biasness is using a stratified sampling.

**(Refer Slide Time 17:14)**

### Stratified samples

- Requires knowledge about distinct, spatially defined sub-populations (spatial subsets such as ecological zones)
- More sample points are chosen in areas where higher variability is expected
- Every unit in a stratum has same chance of being selected
- Using same sampling fraction for all strata ensures proportionate representation in the sample
- Disadvantage: sampling frame of entire population has to be prepared separately for each stratum, it can require large sample



So but it is very important to understand it requires a knowledge of distinct spatially defined subpopulations in that entire dataset. So you should know in the entire dataset what are the distinct sub groups that are there? And how distinct they are and what kind of knowledge you need? If you do not understand that then you cannot do a stratified sampling instead this rather better to do a strategic sampling.

More sample points are chosen in the areas where higher variability is expected. So for example they in the rainfall so the variability is much you will look at more sampling point. Where the variability is less you will look at the lesser sampling points. So every unit in this stratum has same chance of being selected using the same sampling fraction for all strata ensures appropriate representation of the sample.

So now that biasness is removed which was there in the in this strategic sampling but in the stratified sampling the main disadvantage is that the sampling frame of the entire population has to be prepared separately for each stratum. That is the time taking task ok. But if done you can get extremely good sampling point all across your region it can require extremely large samples that is what the stratified sampling explains. It needs very big sampling datasets. So always if someone is interested in extremely clean data extremely the data that has no biased then use stratified sampling which would actually give you extremely good results.

**(Refer Slide Time 19:12)**

## Secondary Data

- More and more ready-made digital GIS data sets become available
- Government agencies: census geography
- Topographic surveys
- Private companies



Then coming to the secondary data. Secondary data is more and more readymade digital GIS data sets it may be through the government agencies like the central, census, geography or geographical department or it may be a topographic surveys that has been done or from the private companies.

**(Refer Slide Time 19:35)**

## Secondary Data

- Meta-data: "data about the data"
  - Procedures used to collect or compile the data
  - Data lineage (Life cycle of data)
  - Accuracy and measurement standards
  - Coding schemes
- Required for both spatial and attribute data

The diagram consists of a central orange circle labeled 'Metadata'. It is surrounded by four other orange circles, each containing a question: 'What?' at the top, 'Where?' on the right, 'How?' at the bottom, and 'When?' on the left. These circles are connected by a circular path.

A photograph of a man in a white shirt speaking, likely the presenter or instructor.

So when you look at the secondary data the first thing that has to be looked at is metadata. As I said previously said metadata is the data about the data ok. So first what is the procedure that this particular data has been used to collect the different procedures whether it has to be used in the form of sampling or what kind of sampling has been used. Then data lineage for example life cycle of that data how it started? How it has been converted into different forms? Now, what is

way of representation that you should know. Then accuracy and the measurements standards this is also extremely important. What kind of accuracy it has I did give you a list of a examples in the tabulated forms? So all those accuracy and the measurements standards has to be looked at. Then the coding schemes, not that every coding schemes is similar. So look at the coding scheme how it has been coded so that is extremely important.

We will also look at some of this coding schemes. So it has to be required for both spatial and attribute data. So metadata is available separately for both spatial and attribute data. So look at both of the meta for look of the so look for both metadata in your secondary data. Without metadata do not use the secondary data, so that would create problems in your data model ok.

**(Refer Slide Time 20:58)**

**Secondary Data**

- Data collected for other purposes can be converted for use in GIS
- Raster secondary data
  - Scanning of maps, aerial photographs, documents, etc
  - Important scanning parameters are spatial and spectral (bit depth) resolution
- Vector secondary data
  - Collection of vector objects from maps, photographs, plans, etc.
  - Digitizing: Manual (table), Heads-up and vectorization
  - Photogrammetry: the science and technology of making measurements from photographs, etc.

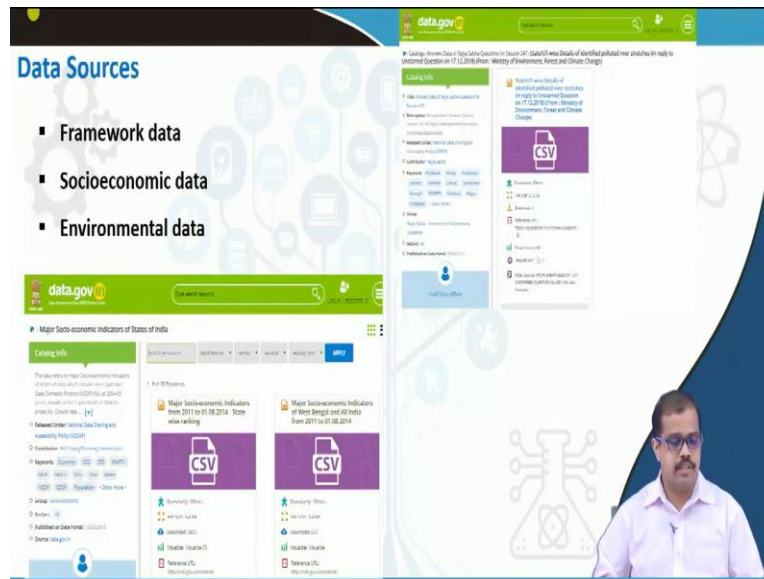
Source: mps.gov

So when we look at the secondary data it can be generated from or converted from various other sources. For example when we look at the raster data model when we look at the secondary data it can be through the scanning of maps, aerial photographs, documents all this becomes a secondary. Any document for example in a government record or a gazetteer set here you can scan it and digitize it in ordered form of secondary data.

Important scanning parameters are spatial and spectral resolutions that you have to look at. Now Vector secondary data is collection of vector objects from the maps or photographs. I will give you some examples of how it is done. Digitizing manually on a table or heads up digitizing and vectorization I will also give you some examples of this in next slides. And photogrammetry the

science and technology of making measurements through a photograph. That also can be a secondary data. So we can have various secondary data that we can generate also can get it from different sources maybe government or private sources for different applications.

**(Refer Slide Time 22:06)**



For example when we look at the data sources we can divide into 3 types one is Frame work data, the second one is Socio-economic data and the third one is an Environmental data. When we look at the framework data so before going into the frame framework data I hope many of you would have look at this particular website which is called data.gov.in. If not, please have a look at this particular data website. It has extremely good data sets or the data has been collected over the period of time.

And this has been made open for the public. There will be certain dataset which are actually outdated but these are also extremely important when if you have certain dataset which are old, so that they give you temporal information. So look at this data.gov.in there is a huge dataset. We also have natural data registry where you can early find out about this data. And also most importantly as whatever the data has been put here as secondary data as metadata.

So whenever you downloading any data from this particular website the metadata associated with it can also be downloaded at least looked at. So please look at this particular website data.gov.in at your leisure. But please look at it is extremely good and very efficient and good effort from government of India in terms of sharing the data to the general public.

**(Refer Slide Time 23:43)**

**Framework Data**

- Reference data to provide context for other data
- Roads, rivers, elevation and contours
- Topographic survey, ordnance survey

The slide features three maps: a red line network map on the left, a topographic map with elevation contours in the center, and a color-coded map on the right. A small inset video of a man in a white shirt and glasses is visible in the bottom right corner of the slide.

So when we look at a framework data as I said it is one kind of data this framework data is actually a reference data. For example, if you look at this entire map if I actually give a reference of may be let us say it a NH and SH are a state highways and national highways. If I give a reference of NH here ok. So that becomes a reference data for any of the other observations. So framework data is basically a reference data that can be used to refer other kinds of data or use it for various analysis as a reference point.

So that is why the framework data is extremely important in order to develop the entire data set. It can be through topographic surveys, it can be through ordnance surveys, it can be elevation dataset, it can be a contours, it can be rivers, it can be roads, any data that you know exact point is a reference data or a framework data or maybe intersection of two roads can be a reference data. So these are certain these are this by framework data extremely important in order to for various analysis or building up the extremely good data model.

**(Refer Slide Time 25:05)**

## Topographic Map

- It is also called as TOPOSHEET
- Scales available- 1:25000, 1:50000, 1:250000
- Prepared by the Survey of India.

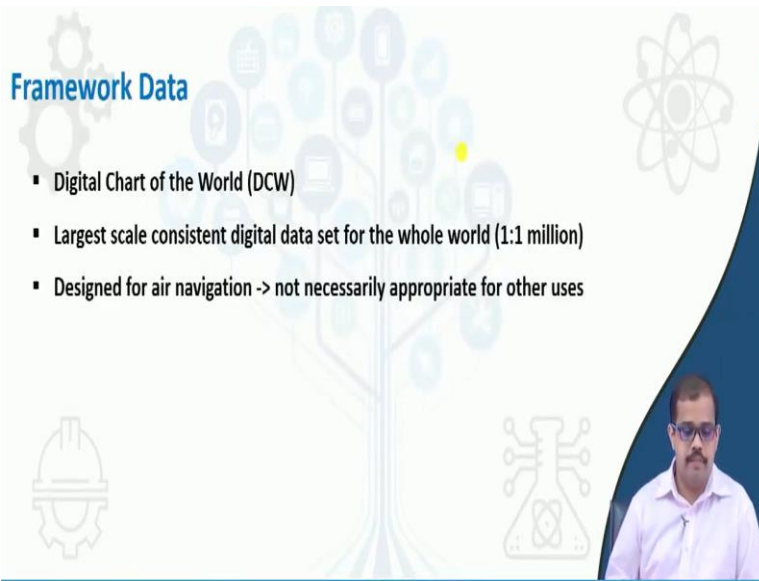


Then you have topographic map ok so it is it is also called as a Toposheet now as of now we have 1 is 50, 1 is to 25000, 1 is to 50000, 1 is to 250000, 1 is to 1 million maps. So these are actually prepared by survey of India. So one of those maps is represented here will look at how this maps are developed. What are different systems of India at different country systems? And what are the properties each and every map has? What each of these represent in a map all of these things will look at when we are studying about a map. So but as of now these are the topography maps also can be used as a framework data. As a source of framework data.

**(Refer Slide Time 25:51)**

## Framework Data

- Digital Chart of the World (DCW)
- Largest scale consistent digital data set for the whole world (1:1 million)
- Designed for air navigation -> not necessarily appropriate for other uses



So it can be framework data the digital chart of the world can be a source of framework data. So largest consistent digital data set for the whole world is 1 is to 1 million. 1 is to 1 million dataset



can is extremely consistent in terms of representing the whole world. So designed mainly for air navigations not necessarily may be appropriate for any other users 1 is to 1 million data.

**(Refer Slide Time 26:22)**

**Socioeconomic Data**

- Data about humans, human activities, and the space and/or structures to conduct human activities
- Demographic data
- Migration
- Housing
- Food
- Transportation
- Economic activity

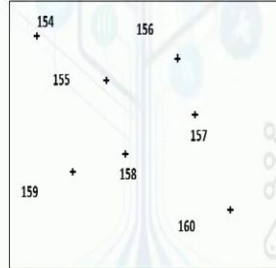
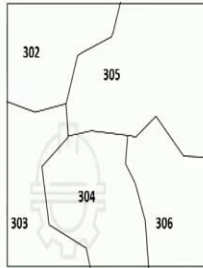
The slide features a central diagram with 'SDG' in a white box. Eight arrows radiate from this box to various icons: a red circle with three people, a yellow bowl of food, a yellow sun, a green circle with a pulse line, a blue circle with a person, a green circle with a globe, a red circle with a book, and a yellow circle with a factory. The background is light blue with faint icons of a gear, a person, a lightbulb, and a network.

The next set of data that you have is a socioeconomic data. So it can be demographic data, so it can be migration data, housing data, food data, transportation data, economic activity data. So these are the data about humans, human activities and the space and structures to conduct human activities. So this becomes extremely important in your database. So when you are collecting the data look at if you have to look at the socioeconomic part of it. So these are certain data sets that you may also have it on the on the government domain that is already available or you may have to collect it as a primary source.

**(Refer Slide Time 26:59)**

## Socioeconomic Data

- Referenced by
  - Administrative units
  - Settlements / villages
  - Individual houses or facilities



So normally social economic data are referenced by as I said administrative units it may be per ward, per district or it may be per village kind of thing. Then it may settlement of villages as I said, then individual houses or facilities if it is a house by surveys it may be individual houses also. So look at what kind of data that you may need in terms of every aspect of collecting the data.

**(Refer Slide Time 27:30)**

## Summary

- GIS Data collection: time consuming, it accounts for 15–50% of total project cost
- Data sources: Digital images, satellite data, GPS data, Survey data etc.
- Primary and Secondary data
- Data sampling methods: Random, systematic and stratified
- Secondary data sources: framework data, socio-economic data, environmental data
- In the next session, we will discuss about data inputs to GIS and more.



So when we have seen this let me summarize this but always remember that there are 2 data sources, the best data source that you can guarantee about to your user is primary data source. When you look at secondary data source populate your database, please look at it metadata.

Without metadata your database is going to be absolutely rubbish. So when you look at we looked at GIS data collection, we also looked at why it is time consuming.

I did explain about it is about 15 to 50% of the total project cost it may involve. Then we looked at data sources like the primary data source and the secondary data source different data sources. I have already spoken about the huge amount of data source in my previous lectures, please look at it. So then we looked at data sampling methods. We have 3 types one is random, systematic and stratified.

Random is used only for the small datasets whereas the systematic sampling can be used for large dataset but would be biased. Stratified is the best way of sampling as of now, but this stratified sampling has huge voluminous work to be done before going into the field. So what kind of sampling is done please look at it. So that gives you a exact way of understanding how the data is generated. How your data is there?

Then the secondary data source then you have framework data, the socio-economic data, you have a environmental data. I have not spoken about environmental data I will give you some examples of environmental data as you progress. So there are huge datasets where we can connect environmental data. I can share it with you over in next classes also. And probably in the next class I will discuss about how you will do a data inputs?

So now we have been speaking about digitizers, we have spoke about the GPS how GPS can be a input as a data we have looked at different ways of inputting the data as plotters, scanners etc., In my next class I would speak about specific ways of how do you input a GIS data. How do you create a data by yourself where if you have a reference data. Thank you very much.