Wendy Thomas

EDDI2015 – Copenhagen

1 December 2015

THE INTERFACE BETWEEN SPATIAL AND STATISTICAL DATA

Abstract

- Information used to create accurate links between data sets at the spatial level
- Structures in commonly used standards to capture this information
- Content that should be provided by data producers in order to support the search for related data in an open-data environment
- Special issues involved in linking historical data and non-standard spatial types

Organization of Workshop

- Who are you and what do you want out of the workshop?
- Differentiating geospatial and statistical data
- Information used to link geography between data sources
  - Statistical to geospatial
  - Statistical to statistical
- How is this linking information captured?
- Special issues of historical data and non-standard spatial areas

Who are you?

- Name and organization
- What do you do with data?
  - Capture/collect?
  - Describe – create metadata?
  - Distribute?
  - Analyze?
- What do you want out of this workshop?

Information used to create accurate links between data sets at the spatial level

What is geospatial data?

- Definition of **Geospatial Data**. **Geospatial data**, GIS **data** or geodata has explicit geographic positioning information included within it, such as a road network from a GIS, or a geo-referenced satellite image. **Geospatial data** may include attribute **data** that describes the features found in the dataset.

What is Geospatial Data? - Research Guides - York University (Sep 14, 2015)
researchguides.library.yorku.ca/content.php?pid=245987&sid=2176375

Define “geospatial data” for a non-GIS professional

answered Aug 5 '10 at 0:36
Don Meltz

- Geospatial data is data that includes location as one of its attributes.
- Not necessarily on the surface of the earth (could be above, as in weather, or below, as in ground water)
- a simple list of counties and their estimated populations is NOT geospatial data, unless it includes the location of each county. E. g., if it includes the state and country the county is in, then it would be geospatial.
- CAD data isn't always, but can be geospatial, if it includes the proper coordinate system/projection.

Geographic Information Systems
http://gis.stackexchange.com/questions/735/define-geospatial-data-for-a-non-gis-professional

Define “geospatial data” for a non-GIS professional

answered Aug 5 '10 at 0:36
Don Meltz

- Geospatial data is data that includes location as one of its attributes.
- Not necessarily on the surface of the earth (could be above, as in weather, or below, as in ground water)
- a simple list of counties and their estimated populations is NOT geospatial data unless it includes the location of each county. E. g., if it includes the state and country the county is in, then it would be geospatial.
- CAD data isn't always, but can be geospatial, if it includes the proper coordinate system/projection.

What I would say is, if it includes enough information for someone to find its location in 3 dimensional space, then it's geospatial data. Specific coordinates (as in GIS data) are not necessarily required. Some might disagree with that, but that's my view. – Don Meltz Aug 5 '10 at 2:09

Geographic Information Systems
http://gis.stackexchange.com/questions/735/define-geospatial-data-for-a-non-gis-professional

What is statistical data?

- **Definition:** Statistical data refers to data from a survey or administrative source used to produce statistics.

Source Publication: Measuring the Non-Observed Economy: A Handbook, OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, 2002, Annex 2, Glossary.

https://stats.oecd.org/glossary/search.asp
OECD Glossary

What is statistics?

- **Statistics** is the study of the collection, analysis, interpretation, presentation, and organization of **data**. In applying **statistics** to, e.g., a scientific, industrial, or societal problem, it is conventional to begin with a **statistical** population or a **statistical** model process to be studied.

Statistics - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Statistics**

Summary

- Because statistical data has a geographic dimension most geographers would think of it as “geospatial data” (where the geographic dimension has statistical attributes associated to it)
- However, statisticians need to have deep and very explicit descriptive information about how and why their data comes into existence (hence DDI) and the ability to transfer data that exhibits known characteristics and definitions (hence SDMX)

So what's the problem?

- Geographers have been very good at describing their "space" and have provided a clear "plug in" point for attaching descriptive data about that space.
- Statisticians assume there is a geographic or spatial dimension to their data but minimize the description of it based on assumptions, lack of understanding, and the use of the spatial dimension as a means of controlling confidentiality.

What's needed

- First a better understanding of what geospatial data is and what it does
- How geospatial data is used to associate "attribute" data to a geospatial location and how different types of geospatial information can be layered to support analysis of interactions between those layers.

Let's start with spatial data

Raster

- In the raster data model, land cover is represented as:
  - single square cells
- Each cell will have a value corresponding to its land cover type.

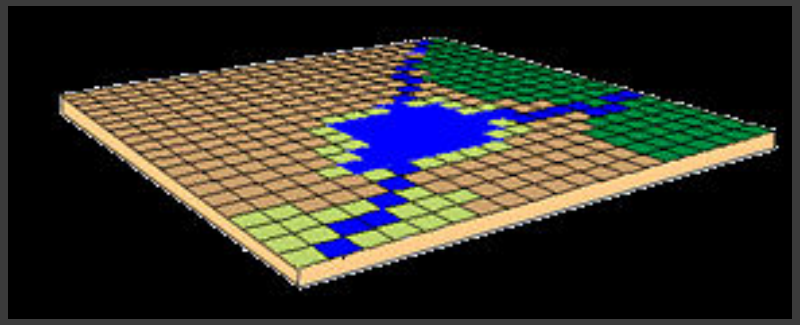
Vector

- In the vector data model, features on the earth are represented as
  - points
  - lines / routes
  - polygons / regions
  - TINs (triangulated irregular networks)

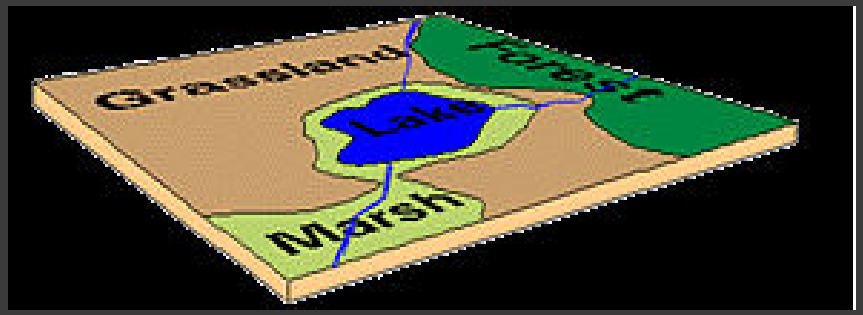
http://gis.washington.edu/phurvitz/professional/SSI/datatype.html

Let’s start with spatial data

Raster



Vector



http://gis.washington.edu/phurvitz/professional/SSI/datatype.html

Let's start with spatial data

Rasters are good at:

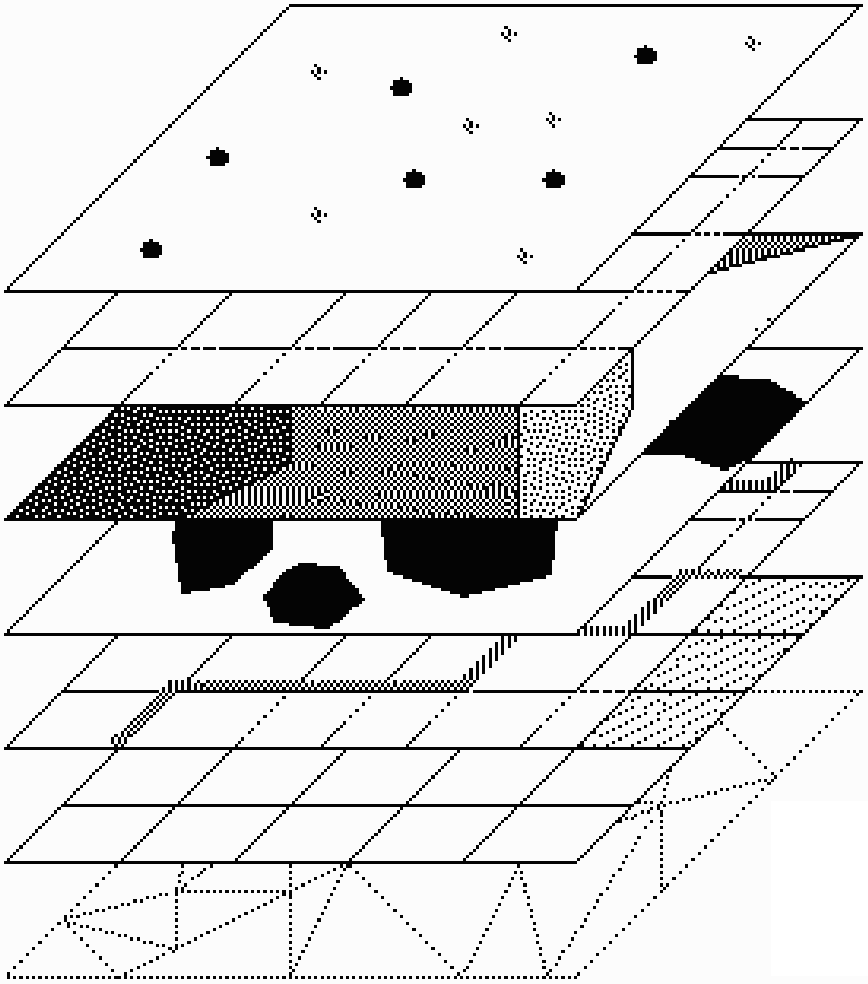
- representing continuous data (e.g., slope, elevation, chemical concentrations)
- representing multiple feature types (e.g., points, lines, and polygons) as single feature types (cells)
- rapid computations ("map algebra") in which raster layers are treated as elements in mathematical expressions
- analysis of multi-layer or multivariate data (e.g., satellite image processing and analysis)
- hogging disk space

Vectors are good at:

- accurately representing true shape and size
- representing non-continuous data (e.g., rivers, political boundaries, road lines, mountain peaks)
- creating aesthetically pleasing maps
- conserving disk space

http://gis.washington.edu/phurvitz/professional/SSI/datatype.html

Storage layers



Feature	File Type
Wells	Point
Streets	Line
Land Use	Polygon
Lakes	Region
Bus Route	Route
Soils	Raster
Elevation	Tin

image @ esri

Associating Attributes

- Attributes are attached to spatial objects i.e.
  - Rastor grid squares
  - Vector
    - Points (single coordinate point)
    - Polygons (a minimum of 4 points where first and last point are the same)
    - Lines (end points linked by a “straight” line)
    - Tins (Triangulated irregular networks)
    - Linear rings (point and radius)
- In order to attach statistical data to a specific spatial object, the data must contain sufficient information to unambiguously identify the associated spatial object.

What does that mean?

- Statistical data sets must either include the specific geographic coordinates needed to fix a location in space, or
- Provide sufficient naming information to associate statistical data to a described location found in a geospatial file by some form of standard naming structure
  - Identify spatial type
  - Naming structure
  - Specific location name

The basics of geospatial description

- Everything is made up of 3 pieces of information:
  - Height (Longitude)
  - Width (Latitude)
  - Depth (Elevation)
- Social science data tends to assume elevation as the surface of the earth unless otherwise specified

Focus on common spatial types in statistical data

- **Point**
  - Defined by 2 (or 3) part number
  - Common case on the surface of the earth: longitude and latitude express as a decimal
  - Common case on an image: Horizontal and vertical offset from a specified location on an image (center, upper left corner, etc.)
  - A point doesn't move, but what it represents might

Examples in statistical data:

- Address
- Centroid of a polygon (i.e. a City, Country, etc.)
- Event location
- Position of an object within a polygon

Focus on common spatial types in statistical data

- **Line**
  - Defined by 2 points and the shortest distance between them
  - Common case on the surface of the earth: River, segment of a polygon boundary
  - Common case on an image: Vein in an MRI
  - A line can change length or direction, the type of shape it represents, and the specific name of the shape

Examples in statistical data:

- Street
- Segment of a polygon (i.e. a City boarder, Country boarder, etc.)
- River
- Route (flight path, transport route)
- Railway line

Focus on common spatial types in statistical data

- **Polygon**
  - Defined by at least 4 ordered points with first and last being the same point (each point pair in the order represents a line)
  - Common case on the surface of the earth: Country boundaries expressed as a series of line segments
  - Common case on an image: Region of a tumor on an MRI
  - A polygon can change shape, the type of shape it represents, and the specific name of the shape

Examples in statistical data:

- Country boundary
- Lake boundary
- Region boundary (geographic or non-geographic)
- Aggregations of smaller polygons (U.S. Metropolitan Areas)

Linking statistical data at the spatial layer

- What type of space is it?
  - Is it a country, region, tract, etc.
  - Is it described in an organized way
    - NUTS system
    - FIPS system (U.S)
- Which space of that type?
  - Unique name
  - Coding system (single string or component parts)
- Do the time frames match?
  - That's the geographic time frame…remember polygons change shape

What type of space is it?

- What type of space is it?
  - Much of our published statistical data is expressed in polygons (Country, State, etc.) either as a level of aggregation or a means of protecting confidentiality
  - For political geography, our spatial types are well organized into structures (hierarchies, etc)
- To link between data sets we need:
  - What structure is being used?
  - What type within that structure is this object?
  - Do I have a means of cross-walking between different structures

Which space of that type?

- Does it have a unique name?
  - Linguistic names are common for higher level geographies
  - For political geography, our location names are well organized into structures that can provide uniqueness.
    - Kansas City, KS is not Kansas City, MO
    - Coding schemes that chain for uniqueness
      - Many states have a county code 053 but 27053 is Hennepin County (053) in Minnesota (27)
- To link between data sets we need:
  - What structure is being used?
  - What name within that structure is this location?
  - Do I have a means of cross-walking between different structures

Do the time frames match?

- Polygons, particularly political boundaries, change shape over time
- Social sciences assumes that the date of the data will match the date of the geography
- So what happens in a time series?
  - The geography changes so that DE is no longer the same shape
  - The data is interpolated to create equivalent shapes over time
- To link between data files
  - Do I explicitly know the GEOGRAPHIC DATE

Summary of linking between statistical data sets

- Spatial type
  - What structure is being used?
  - What type within that structure is this object?
  - Do I have a means of cross-walking between different structures
- Location
  - What structure is being used?
  - What name within that structure is this location?
  - Do I have a means of cross-walking between different structures
- Time
  - Explicit geographic date

Linking statistical data to geospatial data

- What is the spatial object type of the data?
  - Point, Polygon, Line, Linear Ring, TIN
  - You can build up from points but can't deconstruct more complex objects
- For Points:
  - Coordinate points
    - Including measurement system and accuracy
- For other objects or points without coordinate information:
  - What type of shape is it?
    - Same as between statistical files
  - Which specific space is it?
    - Same as between statistical files
  - Do the time frames match?
    - Same as between statistical files

Structures in commonly used standards to capture this information

DDI Codebook

- DDI-C standCtgry
  - Standard category codes used in the variable, like industry codes, employment codes, or social class codes. The attribute "date" is provided to indicate the version of the code in place at the time of the study. The attribute "URI" is provided to indicate a URN or URL that can be used to obtain an electronic list of the category codes.
  - Can be used to reference a geographic structure or location coding system as well as provide the geographic date

DDI-Codebook (cont.)

- Categories or text
  - Geographic structure codes are often entered in the documentation with reference to the standard code
  - Longer lists are often listed as text strings as they only have meaning as part of a concatenated code (State-County-Tract)
- The attribute "geoVocab" records the coding scheme used in the variable.
- Note: Nothing prevents you from using both text representation and standard categories representations which includes the ability to specify a geographic date

DDI-Codebook (cont.)

- geoMap
  - This element is used to point, using a "URI" attribute, to an external map that displays the geography in question. The "levelno" attribute indicates the level of the geographic hierarchy relayed in the map. The "mapformat" attribute indicates the format of the map.
- The "geog" attribute indicates whether the variable relays geographic information.

DDI-C: Outstanding Issues

- The only means of exposing location names (linguistic or code) outside of the data is through Standard Category
  - No means of limiting the extent of the list used
  - No means of describing how to construct a unique code for a specified geographic structure

DDI-Lifecycle

- Spatial coverage provides complete high level information to specify both general information and expose detail found in the data file including:
  - Spatial object
  - Geography Structure Variable reference
  - Geographic Structure reference
  - Geographic Location reference
  - Top Level reference
  - Lowest Level reference

GeoSample2.xml

DDI-Lifecycle (cont.)

- Geographic Structure allows complete profiling of very complex structures
- Geographic Location allows complete profiling of locations (including valid dates), linked to the geographic structure and to external shape files
- Both Geographic Structure and Geographic Location can be used as a representation (in whole or part)

MPCgeo.xml

DDI-Lifecycle (cont.)

- Both Geographic Structure and Geographic Location can be used as a representation (in whole or part)
  - It can specify how to parse the full identifier (i.e. if the Variable contains the County code the Representation can specify that the County code is 3 characters beginning at character 3 of the unique code)
- A Variable can represent a constructed string composed of multiple individual variables (State, County)

RepresentationExamples.xml

SDMX

- SDMX assigns “roles” to Concepts used to define the Data Structure Definitions (DSD)
- REF_AREA or COUNTERPART_AREA are two DSD’s that reflect geography
  - These classifications currently focus on higher level geography resulting in ambiguity for smaller geographic areas within and between countries. In addition the classifications do not all carry clear valid date ranges for the related “geographic footprint”, the actual boundaries of the location at a given point in time.
- The FDI_IO-Data_Request_Codification_v1.0.xlsx contains the Eurostat FDI Geographical classification (Eurostat FDI Geo classification) and the OECD FDI geographical classification (OECD FDI GEO classification)
  - These provide the 2-character ISO code and name for countries and regions (or sets of countries) but provide no linkage to spatial representations
- http://sdmx.org (search fdi_io_data_request_codification)

GSBPM/GSIM/CSPA

- There is a link between the Data Set in GSIM and the Population that contains information on “geographic boundaries”.
- The Population can be associated with any Dimension or Attribute reported in a Data Set.
- Therefore, “geography” is supported conceptually in GSIM but GSIM does not elaborate on how this information can be used.
- Geography can also be described as a statistical classification with specification of the locations and description of the levels providing the geographic structure

Content that should be provided by data producers in order to support the search for related data in an open-data environment

Study level information

- Bounding Box
- Location names/codes (specified vocabulary)
- Geographic date
- Spatial object
- Geographic structures identified (specified vocabulary)
- Top level and lowest level structures
- General statement of spatial coverage

How do geographers search?

- Bounding Box
- Spatial objects
- Using standard names
  - ISO codes (country and regional geography)
  - Country specific codes (internal geography)
  - Using non-linguistic codes
- Top level and lowest level objects

Inbox - wlt@umn.edu - U | DataONE Data Catalog

https://search.dataone.org/#data/page/0

DataONE

About News Participate Resources Education Data

DATA CATALOG: Search Summary Jump to: DOI or ID Go

Search

Search phrase

Filter by:

- Data attribute
- Data files
- Member Node
- Creator
- Year
- Identifier
- Taxon
- Location

Datasets 1 to 25 of 169538

1 2 3 ... 6782 Next

Sort by Most recent

Koch, Katrin, Algar, Dave, Searle, Jeremy B., Pfenninger, Markus, and Schwenk, Klaus. 2015. **Data from: A voyage to Terra Australis: human-mediated dispersal of cats.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.6t066?ver=2015-11-30T16:25:15.770-05:00.

Scott Collins. 2014. **Monsoon Rainfall Manipulation Experiment (MRME) Meteorology Data from a Chihuahuan Desert Grassland at the Sevilleta National Wildlife Refuge, New Mexico (2010- present).** U.S. LTER Network. https://pasta.lternet.edu/package/metadata/eml/knb-lter-sev/301/1.

Macagno, Anna L M, Moczek, Armin P, and Pizzo, Astrid. 2015. **Data from: Rapid divergence of nesting depth and digging appendages among tunneling dung beetle populations and species.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pc66n?ver=2015-11-30T12:07:04.856-05:00

Hide Map »

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. 1312 Basehart SE (MSC04-2815; 1 University of New Mexico Albuquerque, NM 87131)

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508 Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

3:12 AM
12/1/2015

Following links

- Clear cross-walks between different vocabularies
- Structured geographic information (as opposed to descriptive narratives) that can provide clear links from statistical data to its associated geography and from that geography to other available data

Outstanding Issues

Issues for Geospatial data

- In general there is great consistency within the geospatial community in terms of structures for describing geospatial data. The inconsistencies lay in the options for geospatial referencing systems and lack of clearly specified cross-walks between these referencing systems whether name or code based.

Issues for Statistical Data

- The primary issues in creating clear linkages between geospatial areas as referenced by statistical data include:
  - Lack of specificity in the statistical data of the geographic referencing system being used
  - Lack of the referential geographic time (often not the same as that of the data)

Community level approach

- Community supported resources providing complete geospatial references (name, code, reference system, layer, and geographic time)
- Expansion of SDMX DSDs to include a full set of information
- Tools to facilitate their use by reference in statistical data
  - Published Geographic Structures and Geographic Locations with selection tools (to provide reference URN
  - Crosswalks between structure and location vocabularies
  - Listing of bounding boxes for common geographies to auto-generate this set of information in a DDI instance

Special issues involved in linking historical data and non-standard spatial types

Historical data

- Lack of shape files
- Inaccuracy and ambiguity in boarders
- Unclear dating
- Uncertainty of the geographic coverage of the data

Non-standard geography

- Lack of shape files
- Ambiguous boarders
- Lack of geographic structure specification
- Unique naming conventions
- Minimum requirements:
  - Define areas by conventional coordinate system in order to place it correctly on the earth's surface
  - Develop and codify the geographic structures and any underlying hierarchies and layers
  - Uniquely identify each area by geographic structure type, name, and geographic date

Links to geospatial metadata information

- Open Geopspatial
  - http://www.opengeospatial.org/standards/wms
- ISO 19115-1:2014 Geographic information - Metadata
  - http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798
- ISO 19119:2005(en) Geographic information – Services
  - https://www.iso.org/obp/ui/#iso:std:iso:19119:ed-1:v1:en
- ISO 19136:2007 Geographic Markup Language (GML)
  - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32554