Aalto University
School of Arts, Design
and Architecture

MAR-E1004 Basics of GIS: Spatial data infrastructure (SDI)

Jaakko Madetoja
25.9.2017
Slides from Salla Multimäki

Learning goals

In this session you will learn

- **To find and use metadata**
- **To be aware of uncertainty**

Why Spatial Data Infrastructure(s)

- **Geographic data resources are valuable**
  - Collecting data is typically costly
  - Some data are impossible to collect without special techniques or without authorities
  - Data are often needed consistently across large areas
    - *Nation-wide, or across international regions,…*

- **Data does not wear out in use, copying data is cheap**
  - So, why not share and reuse it

- **We need to know what data resources are available**
  - What kinds of data, are they fit for my purpose?

… and to get them in use

What is SDI?

“An SDI is a coordinated series of agreements on technology standards, institutional arrangements, and policies that enable the discovery and use of geospatial information by users and for purposes other than those it was created for.“ (Kuhn, W. 2005)

Metadata – description of data

- **“data about data”, or rather “information about data”**

Metadata standard defines the structure and elements of description

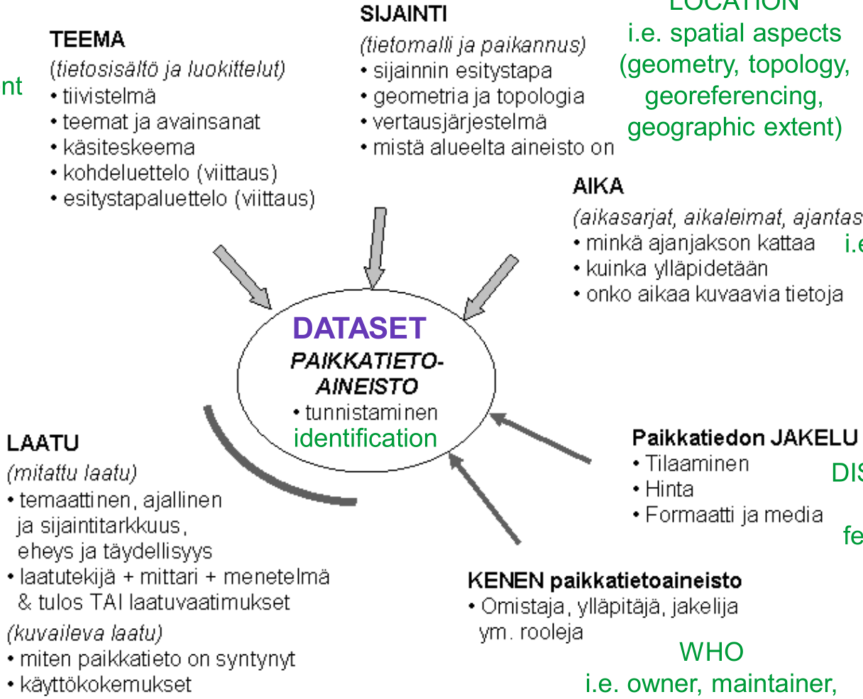
⇒ Standard descriptions allow comparison of datasets by metadata

- **JHS 158 (2005/2012) Paikkatiedon metatiedot (Metadata for the Geographic Information) JHS julkisen hallinnon suositus (Recommendation for the public administration) in Finnish, but with English terms of ISO 19115 in annexes 1-3 (see Annex 2)**

- **ISO 19115-1:2014 Geographic Information – Metadata – Part 1:**
- **Fundamentals**
  ISO 19115-2:2009 Extensions for imagery and gridded data
  ISO 19131:2007 Data product specifications
- **INSPIRE**
  Implementation rules for metadata (~subset of ISO 19115)

What we need to know about data

Framework of metadata content



THEME
i.e. data content
(abstract,
keywords,
object
catalogue)

TEEMA
(tietosisältö ja luokittelut)
- tiivistelmä
- teemat ja avainsanat
- käsiteskeema
- kohdeluettelo (viittaus)
- esitystapaluettelo (viittaus)

SIJAINTI
(tietomalli ja paikannus)
- sijainnin esitystapa
- geometria ja topologia
- vertausjärjestelmä
- mistä alueelta aineisto on

LOCATION
i.e. spatial aspects
(geometry, topology,
georeferencing,
geographic extent)

AIKA
(aikasarjat, aikaleimat, ajantasaisuus)
- minkä ajanjakson kattaa
- kuinka ylläpidetään
- onko aikaa kuvaavia tietoja

TIME
i.e. temporal aspects
(temporal extent,
updating)

DATASET
PAIKKATIETO-AINEISTO
- tunnistaminen

identification

QUALITY
(qualitative
and
descriptive
quality
elements)

LAATU
(mitattu laatu)
- temaattinen, ajallinen ja sijaintitarkkuus, eheys ja täydellisyys
- laatutekijä + mittari + menetelmä & tulos TAI laatuvaatimukset

(kuvaileva laatu)
- miten paikkatieto on syntynyt
- käyttökokemukset

Paikkatiedon JAKELU
- Tilaaminen
- Hinta
- Formaatti ja media

DISTRIBUTION
(access,
fees, formats)

KENEN paikkatietoaineisto
- Omistaja, ylläpitäjä, jakelija ym. rooleja

WHO
i.e. owner, maintainer,
distributor etc.

What we need to know about data

Fitness of datasets for integration

- **Consequences of geometric modelling**
  - e.g. gridded data (resolution & origin and orientation) & vector data
  - e.g. measurement points, sample points – effects of interpolation
  - Notice: the coordinate reference systems

- **Consistency on the level of detail ~ scale, generalization of datasets**
  - do datasets fit geographically

- **Semantic consistency (cf. conceptual model)**
  - object classes, classifications of attributes/variable
  - enumeration units

- **Consistency in quality**
  - Up-to-dateness, or rather the consistency of dates, e.g. when integrating real-time data with static framework data (such as, traffic data for real-time navigation & road network: construction and maintenance work)

A look at different services

- **Paituli (https://avaa.tdata.fi/web/paituli)**
    - Lots of data sets from different institutes

- **Paikkatietoikkuna (https://kartta.paikkatietoikkuna.fi/)**
    - Good for viewing data

- **Paavo (http://www.stat.fi/tup/paavo/index_en.html)**
    - Socio-economical data from postal code areas; see also other data sets from Statistics Finland

- **SYKE (http://www.syke.fi/en-US/Open_information/Spatial_datasets)**
    - Environmental data sets

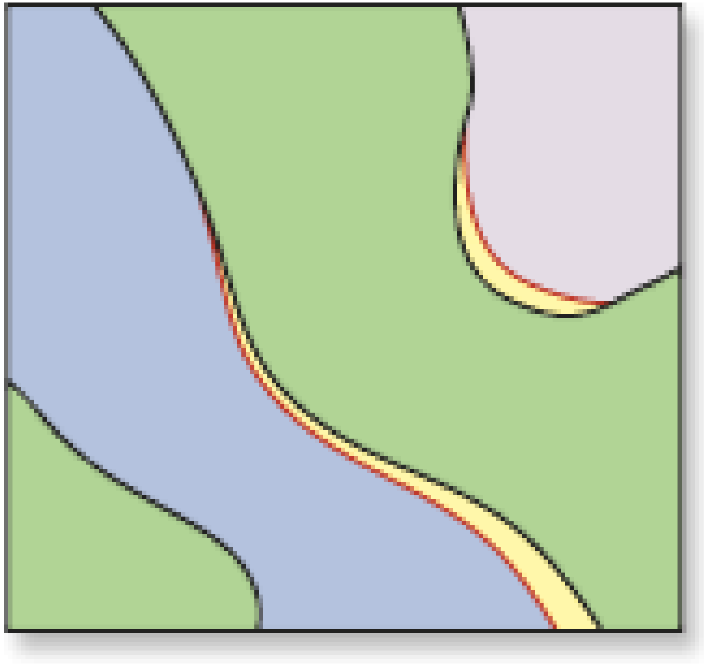
- **OpenStreetMap (https://www.openstreetmap.org)**
    - Open data and map created by users; pros and cons of volunteered information

Uncertainty in geoinformatics

- **"Geographic data is a representation of the real world"**
  - However, it captures only a very limited fraction of the complex and infinitely detailed variation of features and characteristics, at a time.
  - Our understanding of the process of real world is limited, and thus the models as well as the methods used in analysis of data in most cases strongly approximate the real processes.
- **It is impossible to make a perfect representation of the world, so uncertainty about it is inevitable**
  - Uncertainty is more than error, for example, inaccuracy, ambiguity, vagueness,…

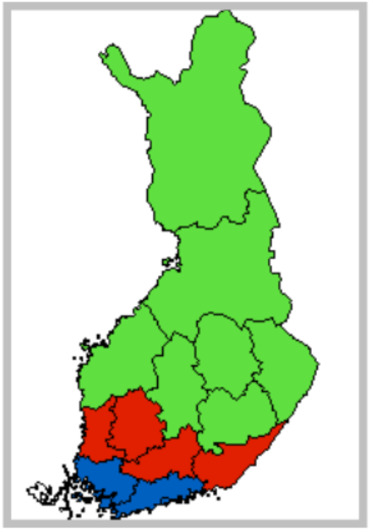
Example: Polygon overlay with erroneous data

Errors cause gaps and slivers

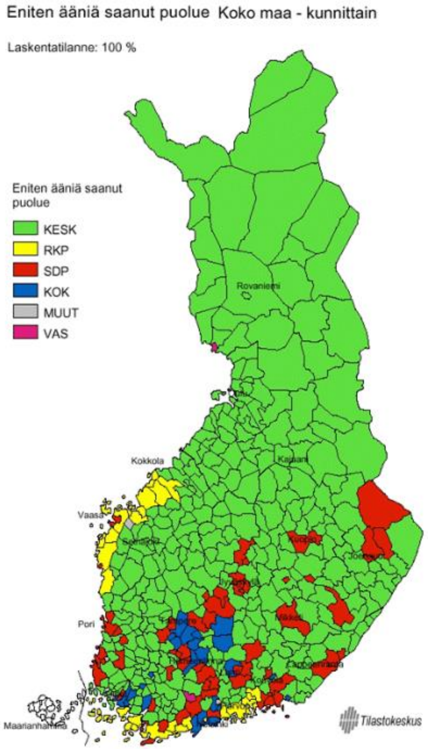


Example: Scale of geographic units matters

Municipality elections 2000

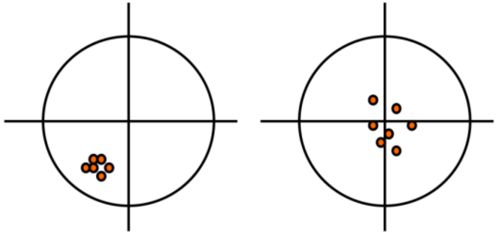


Area units make a difference!



Measurements of uncertainty: Interval/Ratio Case

- **Errors distort measurements by small amounts**
- **Accuracy**
  - refers to the amount of distortion from the true value
- **Precision**
  - refers to the variation among repeated measurements
  - also refers to the number of significant digits (in reporting)



Measuring Error

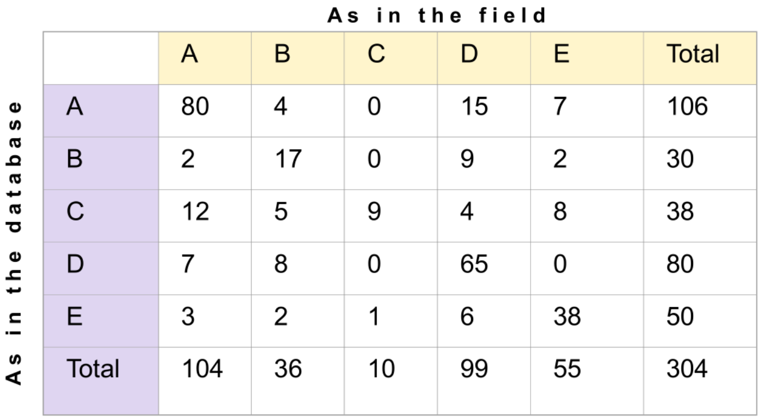
- **Root Mean Square Error**
  - that is, the square root of the average squared error
  - The primary measure of accuracy in map accuracy standards and GIS databases
    - *E.g. elevations in a digital elevation model might have an RMSE of 2 m*
  - The abundances of errors of different magnitudes often closely follow a Gaussian or normal distribution
    - *The range between –1 standard deviation and +1 standard deviation encloses 68 % of the area under the Gaussian curve, indicating that 68 % of observations will fall between these limits.*
    - *2,5 x standard deviation for 95 % reliability*

Statistical measures of uncertainty: nominal case

Misclassification matrix (confusion matrix):

Measure of the accuracy of nominal attributes

- compares recorded classes (the observations) with classes obtained by some more accurate process, or from a more accurate source (the reference)



As in the field

As in the database

	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	0	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

Uncertainty in the analysis

- **The effects of errors and uncertainty on the analysis results**
  - Almost every input to a GIS is subject to error and uncertainty
    - *Location measuring, concept model, rounding, sampling, interpolation, …*
  - In principle, every output (= analysis result) should have confidence limits or some other expression of uncertainty

- **Error propagation (virheen kasautuminen): given the error in the input, what is the variation of output?**
  - Generally done using Monte Carlo simulation: 1) Generate a realization for the input using information about errors 2) Calculate output for it 3) Repeat many times

Living with Uncertainty

- **Data obtained from others should never be taken as truth**
  - efforts should be made to determine quality
- **Effects on GIS outputs can be much greater than expected**
  - there is an tendency to regard outputs from a computer as the truth..
- **Use as many sources of data as possible**
  - and cross-check them for accuracy
- **Be honest and informative in reporting results**
  - add plenty of caveats and cautions

Further reading for uncertainty

Longley et al. (2011) Geographical information systems and science (e-book)

Ch. 6 Uncertainty

Homework 2

The Data for session 3 in MyCourses includes data set CORINE. Your task is to find out *what the data is about*, *what the different values mean* and *visualize* (i.e. create a simple map) *the layer in ArcMap using ”logical” colors* that fit to the theme of the data. You have to decide the attribute that you want to visualize. Report your results for all these aspects.

Hints for visualization: You need to decide between using

- **a single symbol for all objects (the default ArcMap option; also from Symbology -> Features -> Single symbol)**
- **different symbols for different categories (Symbology -> Categories -> Unique values; choose Value Field and click Add All Values to get a list of all values)**
- **graduated colors for continuous values (Symbology -> Quantities -> Graduated colors; choose the attribute for visualization using Value)**

For more information, see ArcMap help in http://desktop.arcgis.com/en/arcmap/latest/map/working-with-layers/a-quick-tour-of-displaying-layers.htm.

Reports to MyCourses as usual.

Only PDF documents, please!