

# **Introduction to spatial information processing**

Ralf Bill (Editor)



**Textbook for international GIS courses**



Internal Reports, Volume 17, 2019  
Rostock University  
Faculty of Agricultural and Environmental Sciences  
Chair of Geodesy and Geoinformatics

---

## Foreword

Maps, which have been a major element of our cultural heritage for centuries, throw light on the distribution of spatial phenomena. Today it is only logical to make the information contained in maps accessible to computers. Transforming this information from its analogue form on paper into a form amenable to computers creates an immense amount of data. It is necessary, therefore, to use special technologies that support the acquisition, management, analysis and visualisation of very large data volumes. Breathtaking advances in computer systems have made it possible to create such technologies, and with more and more computer power available for less and less money, the use of this technology is becoming interesting and affordable for a growing number of users.

This textbook is intended to provide an introduction to these **spatial information processing technologies**. It includes chapters on geographic information systems (GIS), Georeferencing and GNSS, Remote Sensing, Cartography, Information systems as well as WebGIS as. In our opinion, the major parts of spatial information processing technologies, often summarized under the term GIS. It was prepared for fundamental courses on GIS and tested in several countries. It is based on the experience of scientists from different disciplines such as cartography, geography, computer science, geodesy, and remote sensing. The explanations and practical tips in this textbook should help, however, to make the first contacts with the subject of computer-based processing of spatial data as rewarding as possible.

Facilities now exist in the form of geographic information systems (GIS) that can be used to process large volumes of spatial data and to present the data in a visible and intelligible form. Geo-information systems are not able to manage and process only (carto)graphic or map data, however, but also attribute data (from such diverse sources as literature references, files, sensors, and, most particularly, corporate and distributed databases) that can be related to this spatially referenced graphic data. It is this capability which makes geo-information systems such an excellent tool for opening up the spatial dimension of reference to data processed by man for complex daily functions in all imaginable areas.

One highly topical example is the ecology sector. The great challenge of the future is how to keep our planet habitable. GIS technology in the form of complex environmental monitoring and information systems can make a small but not insignificant contribution in this respect. Geo-information systems can serve as an instrument for studying and analysing relationships in our ecosystem, be it for the diverse spatial and thematic dimensions existing in companies, or for local authorities, entire countries, or the whole world. Google Earth and other Earth Explorers are demonstrating this in a very easy-to-use way.

These spatial information processing technologies are going to produce some lasting changes, particularly in the business world. Whether customers, suppliers, competitors, prices or sales figures: for each item of data in business life it is possible to create a spatial connection. Geo-information systems will help their users interpret this spatial dimension and reach decision and conclusion of a new quality. Systems of this type are finding great acceptance in various fields of professional activity since beginning of the nineties. Users range from banks, insurances, real estate businesses, construction firms and transport companies via local and governmental agencies to citizens using web-based geodata portals or location-based services on mobile phones. Geo-information systems represent an immense growth market for the years ahead, especially in the developing countries.

The necessary technologies are available and display a high standard of performance. Powerful hardware, standards in software, the merging of functions from the graphic data processing area and data base management systems and service-based technologies in the internet have all made their contributions and facilitate close interaction.

Spatial information processing technologies are of increasing importance worldwide. They support spatial planning, environmental management, utility administration, land registration, etc. There is a need for a fundamental education and training in GIS for all spatially-related study programmes worldwide, and especially in developing countries.

Ralf Bill, Editor

August 2019



---

## Table of contents

Foreword.....	1
Table of contents.....	3
<b>Part A Geographic Information Systems (GIS).....</b>	<b>9</b>
<b>1 Introduction and definitions.....</b>	<b>11</b>
1.1 Basic terms and definitions.....	11
1.1.1 Basic components of an information system.....	11
1.1.2 Geo-information systems (GIS).....	12
1.1.3 Historical development.....	12
1.2 One world – many views.....	12
1.3 Primary and secondary spatial metrics.....	13
<b>2 Data modelling in a GIS.....</b>	<b>14</b>
2.1 Data types.....	14
2.1.1 Geometry versus topology, vector versus raster data.....	14
2.1.2 Attribute data.....	15
2.1.3 Graphical description.....	15
2.2 Thematic modelling.....	15
2.3 Data and process modelling.....	17
<b>3 Functional components of a GIS (IMAP).....</b>	<b>18</b>
3.1 Data input.....	18
3.2 Data modelling and management.....	18
3.3 Data analysis.....	19
3.3.1 Geometrical methods.....	20
3.3.2 Topological methods.....	22
3.3.3 Temporal methods.....	23
3.3.4 Set methods.....	23
3.3.5 Statistical methods.....	24
3.3.6 Models and simulations.....	25
3.4 Data Presentation.....	25
<b>4 Fields of application and markets.....</b>	<b>26</b>
4.1 Land information systems (LIS).....	26
4.2 Network information systems (NIS).....	26
4.3 Regional planning or statistical information systems (RIS).....	27
4.4 Environmental information systems (EIS).....	27
4.5 Specialist information systems (SIS).....	27
<b>5 GIS products.....</b>	<b>27</b>
5.1 Hardware, software and data.....	28
5.2 GIS product categories.....	28
5.3 Commercial versus open source GIS products.....	28
5.3.1 ArcGIS.....	29
5.3.2 QGIS.....	29
<b>6 Trends and outlook.....</b>	<b>29</b>
<b>7 Summary.....</b>	<b>30</b>
<b>References.....</b>	<b>31</b>
Textbooks.....	31
Online resources.....	31

<b>Part B Georeferencing and GNSS .....</b>	<b>33</b>
<b>1 Introduction .....</b>	<b>35</b>
1.1 Geodesy and its relation to other disciplines and sciences.....	35
1.2 The profession and practice of geodesy.....	36
<b>2 Physical, mathematical and geometrical fundamentals.....</b>	<b>36</b>
2.1 Introduction.....	36
2.2 The shape of the Earth.....	36
2.3 The geoid and vertical reference systems.....	37
2.3.1 The geoid.....	37
2.3.2 Vertical reference systems for height measurements .....	38
2.4 The reference ellipsoid.....	39
2.5 Datum.....	39
2.6 Coordinate systems .....	39
<b>3 Coordinate reference systems (CRS).....</b>	<b>40</b>
3.1 Introduction.....	40
3.2 Geodetic coordinate reference systems (Geodetic CRS).....	41
3.3 Projected coordinate systems (PCS).....	42
3.4 Map projection.....	43
3.4.1 Transverse Mercator (Gauss-Kruger) and Universal Transverse Mercator (UTM) .....	44
3.5 Metadata descriptions of the spatial reference system.....	46
3.5.1 EPSG Geodetic Parameter Dataset (EPSG Dataset).....	46
3.5.2 Well-known text compliant with ISO 19162 .....	47
3.5.3 Others.....	47
<b>4 Coordinate operations.....</b>	<b>48</b>
4.1 Definition: conversion versus transformation .....	48
4.2 Methods of coordinate transformations between CRSs.....	49
4.3 Practical notes.....	51
<b>5 Surveying with GNSS.....</b>	<b>51</b>
5.1 GNSS introduction.....	51
5.2 Introduction to NAVSTAR GPS.....	52
5.3 How GPS works.....	53
5.3.1 Ranging Measurements.....	53
5.3.2 Satellite positioning.....	55
5.3.3 Satellite trilateration.....	55
5.3.4 Accurate timing.....	56
5.3.5 Correcting errors with broadcasting parameters for global correction models.....	56
5.3.6 GPS for absolute positioning.....	56
5.4 Restriction of accuracy and techniques to avoid this .....	57
5.5 Measurement methods.....	58
5.5.1 Meter- and sub-meter accuracies.....	58
5.5.2 Centimeter accuracies .....	58
5.5.3 Post-processing .....	58
5.6 Accuracy management.....	59
5.6.1 Quality of satellite distribution.....	59
5.6.2 Quality of satellite signal reception.....	60
5.6.3 Quality of differential correction signal reception .....	60
5.7 GNSS augmentation systems.....	60
5.7.1 Satellite-based augmentation system (SBAS).....	60
5.7.2 Ground-based augmentation system and ground-based regional augmentation system.....	62
5.8 Comparison of the GNSS.....	62
5.8.1 The renewed GPS.....	62
5.8.2 GLONASS .....	62
5.8.3 Galileo.....	63
5.8.4 BeiDou 2 .....	63
5.8.5 Summary of GNSS Signals.....	64

<b>6</b>	<b>Position accuracy measures .....</b>	<b>65</b>
6.1	Definitions .....	65
6.2	Horizontal position accuracy measures.....	66
6.2.1	Precision indexes and probability levels .....	66
6.2.2	Horizontal position circular precision indexes .....	66
6.2.3	Root mean square (RMS).....	66
6.2.4	CEP (CEP 50) .....	67
<b>7</b>	<b>Mobile GIS .....</b>	<b>67</b>
7.1	Introduction.....	67
7.2	Hardware, components and technologies.....	67
7.3	Operating software.....	68
7.4	Mobile GIS software and geo-apps.....	69
7.5	Measurement-based GIS .....	69
<b>8</b>	<b>Digitising and georeferencing .....</b>	<b>69</b>
8.1	Scanning.....	69
8.1.1	Resolution .....	70
8.1.2	Colour depth and file size.....	70
8.1.3	File formats and compression.....	71
8.2	Georeferencing .....	71
8.2.1	Residual and root mean square (RMS).....	73
8.3	Digitising .....	74
	<b>References .....</b>	<b>74</b>
	<b>Part C Remote Sensing .....</b>	<b>77</b>
<b>1</b>	<b>Introduction .....</b>	<b>79</b>
1.1	Why remote sensing? .....	79
1.2	What is remote sensing? .....	79
1.2.1	Definitions.....	79
1.3	History of remote sensing .....	80
<b>2</b>	<b>Physical basics of remote sensing .....</b>	<b>80</b>
2.1	Light, atmosphere and reflection properties of the Earth surface.....	80
2.1.1	Atmospheric correction .....	82
2.1.2	Remote sensor resolution .....	83
2.2	Spectral signatures .....	83
2.2.1	Indices .....	84
<b>3</b>	<b>Sensor systems.....</b>	<b>85</b>
3.1	Earth observation satellite systems.....	85
3.1.1	Landsat .....	85
3.1.2	Sentinel.....	86
3.1.3	High spatial and temporal resolution Earth observation satellites.....	88
3.1.4	Availability and Prices .....	90
3.1.5	Digital airborne frame camera.....	90
3.1.6	UAS.....	91
3.1.7	Flight planning .....	93
3.1.8	UAS Photogrammetry .....	94
<b>4</b>	<b>Remote sensing applications .....</b>	<b>96</b>
<b>5</b>	<b>Basics of digital image processing / image analysis .....</b>	<b>97</b>
5.1	Filtering.....	97
5.2	Pre-processing operations.....	97
5.3	Multispectral classification.....	97
5.3.1	Multi spectral land use classification .....	99

5.4	Digital change detection.....	100
<b>6</b>	<b>Recent developments and research issues .....</b>	<b>102</b>
	<b>References .....</b>	<b>102</b>
<b>Part D Cartography and Mapping .....</b>		<b>103</b>
<b>1</b>	<b>Introduction .....</b>	<b>105</b>
1.1	What is cartography? .....	105
1.2	Map use .....	106
<b>2</b>	<b>First steps to create a map .....</b>	<b>106</b>
2.1	Map scale .....	106
2.2	Generalisation.....	107
<b>3</b>	<b>Symbolisation .....</b>	<b>109</b>
3.1	Graphic variables .....	109
3.2	Symbol scale.....	110
3.3	Presentation of relief and terrain.....	112
<b>4</b>	<b>Labelling .....</b>	<b>114</b>
4.1	Map fonts .....	114
4.2	Legend design .....	116
4.3	Map layout.....	116
<b>5</b>	<b>Topographic maps .....</b>	<b>117</b>
5.1	Introduction.....	117
	Germany.....	118
<b>6</b>	<b>Thematic maps.....</b>	<b>120</b>
6.1	Introduction.....	120
6.2	Types of data.....	120
6.3	Classification methods .....	121
6.4	Presentation methods.....	125
<b>7</b>	<b>Map-related representations .....</b>	<b>129</b>
7.1	Introduction.....	129
7.2	Special types of topographic maps.....	130
<b>8</b>	<b>Digital cartography.....</b>	<b>131</b>
8.1	Introduction.....	131
8.2	Colour models.....	132
8.3	Raster and vector data in digital mapping.....	134
8.4	Raster data formats in digital mapping .....	136
8.5	Digital data capture.....	138
8.6	Multimedia maps.....	140
8.7	3D visualisations .....	141
	<b>References .....</b>	<b>143</b>
<b>Part E Information Systems and Databases .....</b>		<b>145</b>
<b>1</b>	<b>Introduction .....</b>	<b>147</b>
<b>2</b>	<b>The relational data model .....</b>	<b>147</b>
2.1	Structure .....	147

2.2	Integrity constraints.....	148
2.3	Operations.....	148
<b>3</b>	<b>Conceptual design.....</b>	<b>149</b>
3.1	Application of a conceptual model in the database design process.....	149
3.2	The entity-relationship model.....	149
3.3	Translation into the relational model.....	150
3.4	Normalisation.....	152
<b>4</b>	<b>Inserts of values.....</b>	<b>153</b>
<b>5</b>	<b>SQL: Structured Query Language.....</b>	<b>153</b>
5.1	First example for an SQL query.....	153
5.2	SQL in detail.....	153
5.3	Ordering and grouping of results.....	154
5.4	Joins in SQL.....	155
<b>6</b>	<b>Updating and deleting values.....</b>	<b>155</b>
<b>7</b>	<b>Characteristics of a database management system.....</b>	<b>155</b>
<b>8</b>	<b>Types of database systems.....</b>	<b>156</b>
	<b>References.....</b>	<b>157</b>
	<b>Part F Advanced Geoinformatics.....</b>	<b>159</b>
<b>1</b>	<b>Introduction.....</b>	<b>161</b>
<b>2</b>	<b>Internet-GIS.....</b>	<b>161</b>
2.1	Introduction.....	161
2.2	Terms and range of applications.....	161
2.3	Internet-GIS technologies.....	162
2.4	Functionality of an Internet-GIS.....	164
<b>3</b>	<b>UMN MapServer.....</b>	<b>166</b>
3.1	MapServer CGI.....	166
3.2	Mapfile.....	167
3.2.1	Map section.....	167
3.2.2	Web section.....	167
3.2.3	Projection section.....	167
3.2.4	Layer section.....	167
3.3	MapServer extensions.....	168
3.4	MapServer Clients.....	169
<b>4</b>	<b>OGC and ISO - interoperability and standardisation initiatives.....</b>	<b>170</b>
4.1	Introduction.....	170
4.2	ISO standards.....	170
4.3	Open Geospatial Consortium standards.....	172
4.3.1	Web Map Service (WMS).....	173
4.3.2	Styled Layer Descriptor (SLD).....	176
4.3.3	Web Feature Service (WFS).....	176
4.3.4	Web Coverage Service (WCS).....	177
4.3.5	Geography Markup Language (GML).....	177
4.4	Metadata standards.....	177
<b>5</b>	<b>Application example: Internet-GIS for municipalities.....</b>	<b>178</b>
5.1.1	Introduction.....	178



---

5.1.2	Data collection and conversion .....	179
5.1.3	System architecture .....	180
5.1.4	Technical realisation and results .....	181
5.1.5	From mapping to GIS.....	182
<b>6</b>	<b>Recent developments and research issues .....</b>	<b>182</b>
<b>6.1</b>	<b>Sensor Web Enablement (SWE) .....</b>	<b>182</b>
<b>6.2</b>	<b>Earth viewers.....</b>	<b>183</b>
6.2.1	Google Earth .....	183
6.2.2	Bing Maps.....	184
6.2.3	NASA World Wind.....	184
6.2.4	Google Maps API.....	184
6.2.5	OpenLayers .....	185
<b>6.3</b>	<b>Mashups .....</b>	<b>185</b>
<b>6.4</b>	<b>GeoRSS .....</b>	<b>186</b>
	<b>References .....</b>	<b>186</b>

**Part A**

**Geographic Information Systems (GIS)**

Prof. Dr.-Ing. Ralf Bill



# 1 Introduction and definitions

Before starting to introduce the basic principles and components of a spatial or geo(graphical) information system we need to define some basic terms. This should help to get a common understanding, especially for users not familiar with information technology and spatial phenomena.

## 1.1 Basic terms and definitions

### 1.1.1 Basic components of an information system

**Data** are simply numbers, characters, images, facts, symbols or other outputs from devices to convert physical quantities into symbols, in a very broad sense. They are somehow related to or describing a feature, an idea, a status or a situation. Such data are typically further processed by a human or input into a computer, stored and processed there, or transmitted (output) to another human or computer (en.wikipedia.org).

**Information** is the result of processing, manipulating and organising data in a way that adds to the knowledge of the receiver. In other words, it is the context in which data is taken. Information is somehow linked to information means, such as a natural language or a computer program. The information content of any message depends on the common understanding of both the sender and the receiver. Information as a concept bears a diversity of meanings or knowledge from everyday usage to technical settings (en.wikipedia.org).

Nevertheless, data in everyday language is often a synonym for information whereas in the exact sciences there is a clear distinction between data and information.

“**Knowledge** is defined (Oxford English Dictionary) variously as (i) facts, information, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject, (ii) what is known in a particular field or in total; facts and information or (iii) awareness or familiarity gained by experience of a fact or situation.” .. “The term knowledge is also used to mean the confident understanding of a subject, with the ability to use it for a specific purpose.” (en.wikipedia.org). Thus, some of the terms we use here may have a completely different meaning in other disciplines and contexts.

**Example:** The following example may illustrate the relations between data, information and knowledge. A given binary digit 11111000, coded in 8 bit, we may call data. With the knowledge how to encode a 8 bit representation – summarising  $0 \times 2^0 + 0 \times 2^1 + 0 \times 2^2 + 1 \times 2^3 + 1 \times 2^4 + 1 \times 2^5 + 1 \times 2^6 + 1 \times 2^7$  – we may easily compute the integer number 248. Setting this into any further context, e.g. querying a database to count the number of parcels in a certain area, we may gain the valuable information, that within a certain area of interest, there exist 248 parcels.

A **system**, a term we also often use in different context, is a set of elements or entities, real or abstract, comprising a whole where each component interacts with or is related to at least one other component and they all serve a common objective (en.wikipedia.org).

**Information systems** are general tools to manage and analyse data. Information systems are based on databases and their database management systems (DBMS). An information system is basically a “question / answer” system for a set of data. System components in information systems are either the information or data sets, which then links to databases or data modelling, or the hardware, software, data and users’ applications which make an information system executable on a computer (Figure 1).

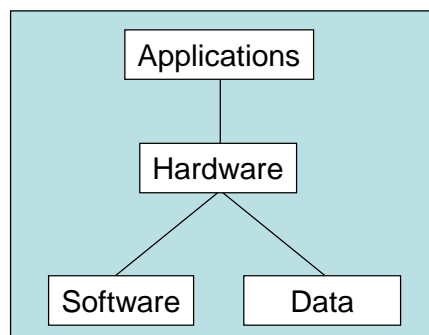


Figure 1: Components of a GIS

Typical examples for information systems may be management information systems, tourism information systems, bank information systems, flight reservation systems or the World Wide Web as one of the largest online information systems available today.

### 1.1.2 Geo-information systems (GIS)

**Definition:** A geo-information system is a computer-aided system consisting of hardware, software, data and applications. It can be used for the digital acquisition, editing, storing, reorganising, modelling and analysis of spatial data, and for presenting this data in alphanumeric and graphic form (short version, translated from Bill, 2016).

ge|o..., Ge|o..., (greek. gē, gaīa), the Greek word for "Earth", underlines that all information is somehow related to the Earth or parts of it, i.e. it is spatially referenced.

Synonyms for these types of systems are also spatial information systems or geographical information systems: we will use the term "geo-information system" (GIS) in this textbook. Beside this textbook there are many other textbooks on the market (Bartelme, 2005, Bernhardsen, 2002, Burrough & Mc'Donnell, Lloyd, 2015, Chrisman, 2001, Demers, 2011, Tomlinson, 2007). Most of the content of this textbook refers to the German textbook of the editor (Bill, 2016). In addition, you may find a lot of relevant content in the internet, for instance in eLearning platforms such as [www.gitta.info](http://www.gitta.info) or [www.opengeoedu.de](http://www.opengeoedu.de).

The four functional components of a GIS are the input, management, analysis and presentation of spatial data, which is often called the **IMAP-model** of a GIS.

In other words, a GIS combines maps and database information. A true GIS links spatial data with geographic information about a particular feature on the map. It is a computer system capable of holding and using data describing places on the Earth's surface. The heart of any GIS is the database (see Part E) through which questions such as what a feature is, where it is, and how it relates to other features can be answered. A GIS gives you the ability to associate information with a feature on a map and to create new relationships that can determine the suitability of various sites for agricultural development, evaluate environmental impact on biodiversity, identify the best location for a new waste disposal site, and so on.

### 1.1.3 Historical development

The advent of the very first computers marked the wish to process spatial data. In the 1960s, however, computers tended to be used for controlling automatic drawing machines that produced maps, CAD drawings and other graphics. The first approaches to spatial information systems began to crystallize a little later out of two separate disciplines. Between 1960 and 1975 geographers as the pioneers began to develop geographic information systems capable of managing nationwide geographic data. In the next step, a number of different nations promoted the development of digital multipurpose land registers called **land information systems (LIS)**. This was the time when national and regional administrations entered into GIS. In spite of dealing with different subjects, the two fields obviously had a number of points in common. Today they can be looked on as subsystems of a GIS, the one with a spatial orientation and the other one more thematically oriented.

The development from **mapping systems**, which are designed for the rather simple and economical production of maps, to today's geo-information systems was completed in the 1980s. This was the time of the companies, using the first real GIS products on the market. At that time many other terms such as mapping systems, computer cartography, **Automated Mapping/Facility Management (AM/FM)** were introduced. Obviously a GIS does far more than a mapping system, i.e. the production of maps (see Part D) is just one of the functions that can be performed by a geo-information system. A further very important area of activity for a GIS is data analysis with presentation of reports, tabular summaries and other forms of data output, all of which serve to aid the decision-making process.

The advent of databases and the increasing number of functions in computer graphics combined with ever more powerful hardware and software pushed the development of geo-information systems, even on desktop computers, at the beginning of the 1990s. When the management of attribute data became possible around the middle of the 1980s through link-ups with relational databases, the bounds of a map's actual contents were well and truly broken.

Today, geo-information systems are being used increasingly as components of comprehensive specialised information systems, e.g. in the environmental monitoring or facility management fields, or even as part of enterprise information systems. Geo-information systems have developed into mature tools whose time is ripe since mid-1990s. Thanks to the availability of powerful computers, modern and intuitive user interfaces, and a large variety of functions, they are able to provide people with extensive help in tackling diverse problems in spatial data processing. Since 2000, with the availability of **standards** and interoperable products, we may see the time of the Open Market, additionally pushed by the availability of **open data**. GIS is a ubiquitous commodity, which is used by us on smartphones, in navigation systems or earth browsers.

## 1.2 One world – many views

The real world, our environment, needs to be considered from various viewpoints, each depending on the observer's subject of concern. In the ideal case – namely that of an integrated GIS environment – the geo-

information system will make a digital record of all the data belonging to a section of the real world. Through the object-related storage of data combined with the integration and digital processing of geometrical primitives, thematic attributes and administrative attributes of the spatial objects, the geo-information system can be used to compile and present the data material according to diverse criteria; the results are an endless variety of models of the real world. The geo-information system gives all users the chance, therefore, of obtaining their own individual view of the stored data.

In Figure 2 we can see various representations – so-called views: such as would be of interest from the viewpoint of the estate register, the utility and tree register, population statistics, land development, traffic route planning and questions of inventory. A view does not always have to be a map-like representation of the data; many groups of users such as tax administration or residents' register also want reports, tables and so on.

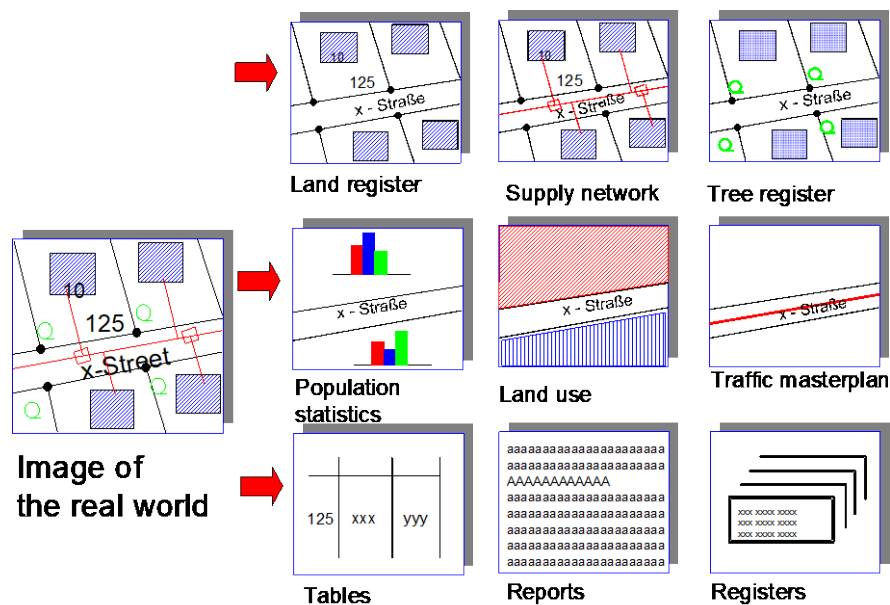


Figure 2: One world – many views

### 1.3 Primary and secondary spatial metrics

To relate features to the Earth we may differentiate between primary and secondary metrics. **Primary metrics**, often also called direct links to the Earth, are two- or three-dimensional coordinates or construction rules. From the latter of these we easily can compute **coordinates**. They have well-defined properties such as a defined metric (see section 3.3.1 for further details), a well-known reference system (see Part B for further details) and high accuracy expectations. From a database perspective it is not easy to find features given their coordinates because this asks for a multidimensional search criteria, which is more complicated than standard queries in databases (see Part E for further details). Many disciplines use coordinates such as engineering in general, surveying, planning and environmental studies.

Beside coordinates, many other users such as tax offices, rescue services or school planning administration use different ways to relate things to the Earth. **Code numbers** such as district numbers, postal codes, area codes etc. define a **secondary metric** with the properties that they are more weakly-defined with low or variable accuracy. For a database they can easily be accessed given a one-dimensional search criteria. Similar to code numbers we may use names of features such as the city name (e.g. Algier, Quito) or the name of a tourist site (e.g. Cat Ba, Vietnam).

Another popular way to organize features in administration and in business-oriented workflows is by using the **address** (city, street name and house number such as Rostock, Justus-von-Liebig-Weg 6). This is used by registration offices, for elections, for logistic purposes, navigation, or emergency services etc. The properties are similar to code numbers and names. It is a weakly-defined metric with low or variable accuracy. To query the database for such features we may need one- to multi-dimensional search criteria: especially a free-text search may cause additional effort for the database.

The important aspect about spatial referencing and metrics is that around 80% of all branch-specific administrative, logistic, and strategic activities in an enterprise or governmental organisation have a spatial reference. Using this allows the more efficient organization of workflows and the linking together of various disciplines such as engineering and planning, administration and government, utility and services, decision makers and citizens.

## 2 Data modelling in a GIS

To explain the working principles and the possibilities of a GIS it is first necessary to discuss the data types and how real world phenomena may be described, modelled and organised in a GIS environment.

### 2.1 Data types

#### 2.1.1 Geometry versus topology, vector versus raster data

An **object** (or **feature**) can be described geometrically by defining it in vector or raster terms. This entails defining the geometry in a uniform reference framework which is usually set by a coordinate system. In a geo-information system, geometry can take the form of a representation based on points, lines or surfaces. While the point- and line-based descriptions produce **vector data**, the surface- or field-based description results in raster data (Figure 3). The point is the carrier of geometric information in the vector data; points are usually described by coordinates. The form of connecting elements can be laid down by additional specifications, e.g. an arc with a radius. The mathematical framework for geometry is given by computational geometry.

**Raster data** are based on picture elements (called pixels) of equal size and uniform distribution (matrix), the geometrical reference being the centre or corner of the element. Lines and surfaces can be thought of as sequences of characteristic points or as neighbouring pixels.






Element	Vector		Raster	
	Digital	Analogue	Digital	Analogue
Point	x,y-coord.	.	Pixels	
Line	x,y-coord.-sequence		Pixels	
Polygon	closed x y coord.-sequence		Pixels	

Figure 3: Vector and raster data – geometric representation of features

Vector and raster data may both be defined in two or more geometric dimensions. Usually, two-dimensional (**2D**) representations using x,y-coordinates are used for thematic mapping in various disciplines. Including the height as an additional dimension allows us to describe the Earth's surface; we call this **2.5D**. It is a very prominent model for environmental issues. For certain applications it may be necessary to model the real world in 3D, or even in **4D** if we include the time as an additional dimension. There are different methods to model **3D**, for instance the 3D line model, the 3D facet model and the 3D volumetric model. Especially Computer-Aided Design (CAD) software deals with 3D volumetric models. The higher the geometric dimension, the more complex the real world may be modelled, but the workload tends to increase dramatically when increasing the dimensionality. Thus, for most practical applications we restrict the data to 2D (a planar view of the earth) or 2.5D (a surface related description of the earth).

These positional data (geometry) must be supplemented by data of the neighbourhood geometry and connectivity (called **topological data**). The topological counterparts of the geometrical elements point, line and surface are called node, edge and mesh. Within topology, the only important fact is that nodes and edges stand in a specific relationship to each other; the geometrical position and form of these relationships is unimportant. The carrier of topological information is the edge. The mathematical framework for topology is given by graph theory and topology.

The easiest way to explain the difference between geometry and topology is to consider the example of a plan for an integrated transport system. The representation of connecting lines in the schema plan is a topological representation in which the only important fact is the existence of connections (edges) between the traffic intersections or stations (nodes): in this representation, the distances and directions between points cannot be measured, only the existence or otherwise of a transport link between them can be determined. A geometrical presentation of the integrated traffic system, on the other hand, shows the transport system overlaid on the town map; distances can be measured and the positional relationships and forms are correctly reflected in accordance with the chosen scale.

Topology must be specified explicitly in the vector world. In the raster world, on the other hand, it is defined by the arrangement of the pixels (rows and columns) and can be illustrated in tree form (e.g. a quadtree). The raster world and the vector world may exist side-by-side in spatial information systems in what is referred to as a hybrid geo-information system. It is possible to use whichever reference system is the most favourable for the particular application. For example, raster data are preferable for continuously-variable area-based phenomena, whereas the

vector world has its strengths in line-based structures. Conversion from vector data to raster data causes relatively little difficulty. Changing from raster data to vector data, on the other hand, still poses a number of problems. Only specific types of maps can be converted economically into vector data by means of raster/vector transformation.

### 2.1.2 Attribute data

A variety of **thematic information** (called attribute data or descriptive data) can be assigned to spatial objects in a GIS. Attributes are all non-geometric elements such as text, numbers, measurements etc. They are captured in a specific context to solve specific problems. **Attributes** occur in analogue form (registers, protocols, notes) as well as in digital form (databases, files). The mathematical basis is set theory and relational algebra. A street name is an attribute of the feature street. For example, the smallest geometrical element of a cadastre or land register is the lot or parcel. A lot may contain certain raw material deposits, so it will also be listed in the resource records. Table 1 shows a collection of attribute data for a lot of land; an object identification code (OID) links this data with the geometrical and topological description of the object in the GIS.

OID	Parcel number	Usage	Area size	Owner name
125	1232-1	Wasteland	2348.25	Berger
126	1342	Building site	519.12	Bill
127	2367-12	Agriculture	1200.76	Nash

Table 1: Attribute data in a real estate/land management application

Attribute data collections are compiled mainly by separate specialised disciplines; in nearly all cases each specialisation collects different data on the object.

### 2.1.3 Graphical description

For mapping the results of GIS work we need additional information. The graphical description is given by graphic data combining the geometric data with graphical elements such as symbols, hatching, grey scales, line sizes, polygon fill etc. This information is often described in analogue form (e.g. in a map legend) telling us how features should be displayed under certain conditions in a thematic map. Usually we include additional text elements for annotation in order to match standard graphic elements. The basic framework for graphical descriptions is defined by cartography, mapping, computer graphics and visualisation (for further reference see Part D).

**Summary:** A geo-information system should be able to store geometric, topologic and thematic data on objects in the real world. It should also be able to supply this data to the various processing routines and to produce graphical representations from that. Taking a street as an example, the street is geometrically defined by a combination of straight lines and curves, it connects a start node with an end node (the topology) and it may carry descriptive information such as the street name, the pavement of the street and the speed restriction. For thematic mapping, the different pavements may be coloured in different colours.

## 2.2 Thematic modelling

Each object is represented in a GIS by a geometrical and topological aspect, possibly a temporal component, and one or several thematic characteristics. A GIS must therefore be able to generate geometrical, topological, temporal and thematic models of the real world and create an image of it in databases. Most of these aspects are standardized, based on the international norm family ISO 191xx on geographic information, from the International Standardisation Organisation (ISO). In this section we will first focus on thematic modelling.

Thematic modelling is understood to be the description, processing and storage of a spatial object's underlying theme. Thematic modelling is always dependent on the particular application, but there are fundamental concepts and hence common features to be found in many tasks of different types. Thematic modelling is used to create an image of specific views – defined by the particular application – of the real world in the GIS database. The system must meet very high demands of flexibility if it is to do justice to the complex relationships of the real world.

This explains the development of thematic models which now find application in all geo-information systems. A term often used in this connection is “**object orientation**”, meaning that the spatial data can be viewed on an object basis. The term conflicts, however, with a current technology in computer science, namely object-oriented programming. Although the latter has certain points in common with the object-based processing of spatial data, it also introduces some completely new concepts such as “encapsulation” and “inheritance”. It is recommended, therefore, to refer to an object-based or **object-related model** whenever the concepts of object-oriented programming are not involved.



Fundamentally, it is possible to distinguish between two different approaches within the thematic modelling field. The older method of separating different themes is based on the layer or foil principle. We therefore call it the **layer model**. The layer carries the geometry of the features and its graphical description, i.e. the style defining how the feature should appear in a graphical form, either in a map or on screen. The layer model is often found in CAD systems. The most recent developments, however, have come up with **object-related models** that are able to reproduce thematic relationships with far more flexibility. These models make use of the object class principle, i.e. they build up a graph which in the simplest case takes on the form of a tree.

An **object class** (or **feature class**) defines the types of objects (or features) that occur in the real world (from the special thematic view). Attributes (for all objects of this class!) characterise the individual object and which expressions (domains) the attributes can adopt. An object class may have an object class code, it has a topologic type (node, edge, mesh), a geometry in vector or raster form, and characteristic attributes (e.g. tree species). For the attributes, domains may be defined (oak, beech, chestnut): the format and status need to be described. At least one form of visualisation is needed. Usually we combine the object class definition in so-called **object class catalogues**, which is a collection of all defined object classes.

**Check:** National mapping agencies usually set up their features of interest for thematic mapping in object class catalogues. You may find it helpful to check if such a catalogue exists in your home country and use it as an example for your own modelling of real world phenomena.

This catalogue may be further structured hierarchically into object groups (e.g. vegetation area) and object domains (e.g. vegetation) and object classes (e.g. crop).

**Example:** Taking a forest as an example for thematic mapping we may describe the properties of the object class forest in the following way. A forest is always a polygonal feature, i.e. its topology is described by a mesh. Forests may carry an attribute which describes what type of forest that particular instance is (i.e. deciduous, coniferous). For any thematic analysis of these attributes we may define how it should be displayed (green, filled). Additionally, for database implementation reasons, we may add an object class code. We may group the object class forest into an object group (e.g. vegetation area) and object domain (e.g. vegetation). The object domain vegetation may contain further object classes beside forest such as agricultural areas.

The individual objects and features occurring in the real world belong to an object class. An object is one out of many similar things belonging to that class. It is a concrete, geometrically delimited feature. In database terms we would call each object an instance of the class. The object class forest carries thousands of individual forests in a country.

Taking another example, the feature or object class “Street” may be defined in the following way: in database terms this will later be called a **relation**.

Street {Topological type, geometry, area size, perimeter, line length,  
street type, surface, speed restriction, direction ...}

In a **relational data model** (see Part E for further details) we would set up a tabular data structure for this. Each individual street (one single feature or object) would then occur as one single row in that table (Table 2).

Street							
Topotype	Geometry	Area	Perimeter	linelength	surface	pavement	speed
Polygon	BLOB	120.534	270.348	125.00	highway	asphalt	120
Polygon	BLOB	272.128	350.543	122.12	municipal	asphalt	60
...	...	...	...	...	....	...	...
Polygon	BLOB	654.584	271.156	257.00	highway	asphalt	100

Table 2: Relational database table for thematic class “Street”

Data modelling remains a complex and problematic area. Different disciplines may define feature classes in different ways: A cadastral surveyor may look at forests only being interested in the area size, the owner and the tree classes, a forest engineer looks into a more detailed classification of the tree types, the age of the trees and the year of afforestation. Thus, we have different meanings based on different attributes we collect. We may find also different geometry for the same feature, either captured by surveying with GNSS resulting in vector data or collected from satellite imagery resulting in raster data.

## 2.3 Data and process modelling

When modelling the real world for GIS usage we have to distinguish two different aspects; we can model either from the information processing perspective (the **data model**) or from the typical workflows (the **process model**) we find in an enterprise or governmental office.

First of all we make an abstraction step. Not all phenomena in the real world are of interest for our task, so we have to define which object classes we recognise, which attributes are of interest to us and which analysis functions and processes will be carried out on this features. This way of modelling has a major impact on the methods and the amount of data that we subsequently capture, how it is represented in the database (storage space and structuring) and which analyses we are able to carry out, i.e. which work flows are we able to support in our company.

In the second step we have to carry out an interpretation related to the way things should be visualised at the end of our processing chain (Figure 4). This defines which end products are we able to produce (maps, reports, animation, data for further use etc.)

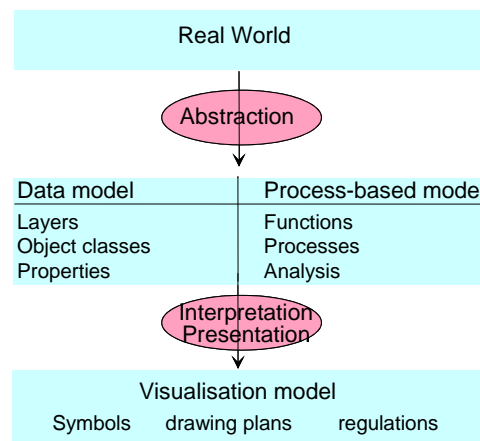


Figure 4: Abstraction and interpretation of real world problems

**Thematic modelling**, described in an earlier section, is a very intensive and very important process a user has to fulfil before starting to use a GIS. In that process it is also necessary to integrate topological, geometrical, and temporal modelling related to the spatio-temporal components of real world phenomena. For the database implementation the user has to know much about logical data models such as relational or object-oriented models.

**Topological modelling** and structuring brings order into such disordered data by reducing the geometrical description to a node/edge model that remains unchanged even if the coordinates are altered. Such a topological structure contains in particular link data and information concerning neighbourhood relationships. In geo-information systems the different forms of topology contribute to solving various problems. They permit, for example, consistency checks and queries as to the shortest routes within spatial data. Functions of this kind are indispensable considering the immense data volume to be dealt with.

**Geometrical modelling** is understood to be the description, processing and storage of the basic geometry of spatial objects using analytical methods and processes of approximation. Geometrical GIS database queries are essential because they form the basis for all spatial questions. Calculations of distance, area and volume are examples. The analysis function of the polygon overlay is based completely on the underlying geometrical model. A geo-information system's flexibility and performance depend, therefore, on how the geometrical modelling is implemented. Many applications use the simple feature model from ISO.

**Temporal modeling** of objects distinguishes between time points and time intervals. It supports the derivation of changes in land use, the tracking of moving objects, the handling of temporal versions of objects or simulations and prognosis.

In the **logical data model** the logical relationships between the various object classes are considered. A distinction is drawn between hierarchical, network, relational and object-oriented data models. The greater the division between the logical data model and the physical data model, the easier it generally is for the user to work with the system and the less the user or subject specialist needs to think about the required details of the low-level implementation.

### 3 Functional components of a GIS (IMAP)

The four main functional elements of a geo-information system have already been mentioned in the foreword. A GIS must be able to input, manage, analyse and present data (**IMAP-model**). The functions of these four basic components will now be described in greater detail.

#### 3.1 Data input

Geo-information systems are based first and foremost on the data with which they work. The digital acquisition of this data is a highly work- and cost-intensive activity; it is also crucial for the use and success of a GIS, particularly in view of the high requirements on the completeness, accuracy and structure of the data base. The choice of data and an adequate acquisition method depends mainly on the application and on the object whose data is to be captured. The framework conditions are set by the available budget and the functions that can be performed by the GIS.

Considering the great variety of possible data – some of which will already be or will become available in digital form as open data – the first step before beginning with data acquisition should be to check to what extent any existing data may be put to use. Nowadays there are more and more stocks of digital data available, leaving users to concentrate on collecting their own new specialised data. Nonetheless, data acquisition represents still a problem for many applications. Before any evaluations can be undertaken in the area of the application, the immense volumes of data from analogue sources need to be converted into a computer-compatible form. This can be done by collecting the data anew from the object itself or its unprocessed image (**primary acquisition**) or by collecting data that already exists in some processed form (**secondary acquisition**).

There are many different methods to capture/acquire data/information in a GIS. We may reference here geodesy and surveying, photogrammetry, remote sensing, digitising or scanning of existing maps and documents, attribute data collection by various methods, integrating existing digital information etc. During this processing step we usually also define the spatial reference systems. For further details see Part B and Part C in this book.

**Check:** Data acquisition should be as accurate and complete as necessary, and it should be as cost efficient as possible. As a rule, it is possible to use more than one method for acquiring the required data; it is advisable, therefore, to weigh the costs, advantages and disadvantages of the various methods and compare them carefully. Before beginning to collect data, you should check exactly if there is any digital data already available elsewhere. Of all the processes in a geo-information system, data acquisition is the most expensive step. As a rule it will cost between 10 and 100 times more than the initial GIS investment in hardware and software.

#### 3.2 Data modelling and management

Each object is represented in a GIS by spatio-temporal aspects together with one or more thematic aspects. Accordingly, a GIS must be able to generate both a space-time and thematic model of the real world and create an image of it in databases. Data modelling relates to **entity-relationship modelling**, relational or object-oriented modelling, standard modelling techniques such as the **Unified Modelling Language (UML)** and databases and information systems, which are covered in more details in part E in this textbook.

Data management and data storage play a central role in a geo-information system. The approach taken in the past was based mainly on management by means of a proprietary file system. Today, the trend in geo-information systems is towards using databases for storage, with the data managed by complex **database management systems (DBMS)**.

The thematically, geometrically, topologically, and temporally structured data in these databases is illustrated with the help of **logical data models** which contain fixed database schemas for implementing the given data structures. The type of data storage on hard disk and the corresponding access mechanism are laid down in the **physical data model**.

A **file system** is characterised by parallel data storage. Users create files (depending on their application), maintain them and use them to solve their own problems. Usually, the data sets thus created are applicable to a unique application only, so the requirements needing to be met by a file system as regards low-redundancy data model, data backup mechanisms and terms of consistency are not very high; indeed, sometimes there may be no requirements at all.

It is precisely these aspects, however, which are of primary concern in database systems. Here the data must be free of redundancies, available for multi-user access in batch and interactive mode, capable of flexible structuring and safe storage, and it must deliver acceptable response times. A plain and simple user interface and a large number of different tools should make the system readily accessible to all users. The only way to meet such demands is with an extensive collection of programs, i.e. a database management system (Figure 5).

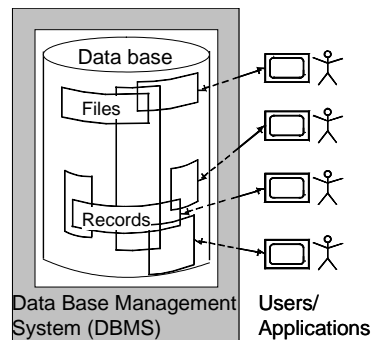


Figure 5: Database management system

If a very high priority is placed on data security, data will represent the biggest part of the investment in a geo-information system. Relational database management systems are now a mature technology and their use is more or less the standard.

### 3.3 Data analysis

The data analysis system is the real performance centre of a GIS. In addition to being the most important component it is also the feature that distinguishes the geo-information system from other systems such as mapping systems and CAD systems, which hardly have any analysis functions. The benefits of a GIS for users depend essentially on whether the system provides them with all the necessary analysis options and on how “easy” it is to use the corresponding tools. This section will now describe the fundamental areas of analysis.

**Definition:** Analysis means the scientific and systematic dissolution and study of problems, an object or correlations, for the division, decomposition of compounds into their components (the opposite of synthesis!). We may differentiate between a qualitative analysis, i.e. according to properties etc., and a quantitative analysis, i.e. according to amount, number, order etc.

The general task of spatial analysis may be seen as follows: given a user-defined task and an information system with information sets/data/observations A, B, C, etc., the task is to establish (a) function(s) through which the available data sets may be involved and manipulated to provide the required output (e.g. presentations such as maps, graphs, reports, ...) related to the problem of the user.

$$\text{Link: } U = f(A, B, C \dots)$$

The functions  $f$  may be typical GIS analysis functions such as selection, Boolean operations, reclassification and polygon overlay with algebraic terms between the data sets being involved.

The analytic tools of a GIS may answer five different types of question:

1. **Location:** What is at a given location? This type of questions seeks to find out what exists at a particular location. A location can be described as a place name, zip code or address.
2. **Condition:** Where does something occur? Using spatial analysis this question seeks to find a location where certain conditions are satisfied (e.g., an unforested section of land at least 2,000 square meters in size, within 100 meters of a road, and with soils suitable for supporting buildings).
3. **Trends:** What has changed since? This question might involve a combination of the first two and seeks to find the differences within an area over time.
4. **Patterns:** What spatial patterns exist? You might ask this question to determine whether cancer is a major cause of death among residents near a nuclear power station. Just as important, you might want to know how many anomalies there are that don't fit the pattern and where they are located.
5. **Modelling:** What if? questions are posed to determine what happens, for example, if a new road is added to a network. Answering this type of question requires geographic as well as other information.

Obviously the term analysis means much more than defined in the beginning of this chapter. To answer the five question categories spatial analysis has to integrate and combine:

- **analysis:** dissecting, decomposing compounds into its components,
- **synthesis:** merging single components to a higher order,
- **simulation:** realistic imitation of technical processes, and
- **prognosis:** assessment in advance (forecast).

Spatial analysis functionality has mathematical foundations in coordinate geometry, numerical methods, topology and graph theory, set theory, relational algebra, statistics and other mathematical and computational disciplines.

In this chapter we describe analysis functions based mainly on their dominating data types, which gives us 4 basic categories:

- geometrical methods: functions related to the geometry data types vector and raster data,
- topological methods: functions dealing with topology and relations between objects
- temporal methods: functions related to the temporal behaviour of objects,
- set methods: functions mainly taking into account the attributes.

In addition we add two more categories; one for statistical methods and one for more complex and general analysis approaches. For further details on spatial analysis we propose “Geospatial Analysis” by de Smith & Goodchild & Longley (2018), which is available online under <https://www.spatialanalysisonline.com/>.

### 3.3.1 Geometrical methods

The term **computational geometry** covers the group of functions which permit measurements and calculations with spatial as well as descriptive data. The main purpose of this function group, which is intensively used in the surveying field in particular, is to derive the numbers and frequencies of spatial and descriptive conditions, distances angles, height differences, areas, volumes, circumferences, etc. from the geometrical data.

Obviously geometrical methods depend on the spatial reference system (see Part B) and geometric assumptions made in mathematics. Therefore we need the definition of distance functions  $d$  and a metric.

**Definition:** Given three points  $P$ ,  $Q$  and  $T$  in a plane we may define: A metric on a set  $X$  is a projection  $d: X \times X$  on  $R_0$  with the following properties for any  $P$ ,  $Q$ ,  $T$  from  $X$ :

- $d(P,Q) = 0$ if $P=Q$	- Idempotence
- $d(P,Q) = d(Q,P)$	- Symmetry
- $d(P,Q) \leq d(P,T) + d(T,Q)$	- Triangle inequality

A pair  $(X,d)$  is called a metric space.

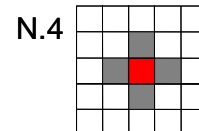
There are different possibilities to describe the distance  $d$  between points  $P$  and  $Q$ . A very common distance function in the vector domain is the Euclidean distance:

**Euclidean Distance:**  $d_E = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$

For raster data it is a little more complicated. Here we have different distance definitions. With  $d_1 = |i - k|$ ,  $d_2 = |j - l|$  for the points  $P(i,j)$ ,  $Q(k,l)$  in pixel coordinates two very prominent distance functions (N.4 and N.8 neighbourhood) can be described.

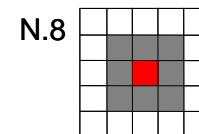
**City-Block-Distance:**

$$d_4 = d_1 + d_2$$



**Chessboard distance:**

$$d_8 = \max(d_1, d_2)$$



**Euclidean Distance:**

$$d_E = \sqrt{d_1^2 + d_2^2}$$

To improve the performance of geometric methods, very often a MER, a **minimum enclosing axis-parallel rectangle** (also called a **bounding box**), is introduced to approximate the complex geometry of lines and polygons. This allows the use of concepts for quick data access and preprocessing and is used, for instance, for the point-in-polygon test and intersection.

**Zone generation** around point, line and area-based objects serves as the basis for calculating the effects of planning measures, for example. A typical question of this type would be: “Which objects (lots, forest, waste disposal sites, etc.) lie within a circumference of 100m around the planned axis of a traffic route?”

**Point-in-polygon**, either with vector data or with raster data, is a standard functionality needed whenever one clicks on a feature at the screen to find out what feature it is.

**Polygon overlay** refers to a method that is capable of creating new data from existing source data by means of geometrical superimposition. It is one of the most important basic operations in a GIS and is needed in all fields of application. In the vector world, polygon overlay runs through the followings steps (Figure 6):

1. Line intersections: Divide all intersecting lines of the starting polygons at their intersections. This results in a list of all nodes and edges, there is no further intersection of polygons.

2. Polygon formation: Link individual edges to form new, closed polygons. The result is a list of all polygons.
3. Overlay identification: Check polygons: which were the original polygons? The attributes are then transferred to the new polygons, either by copying or linking.

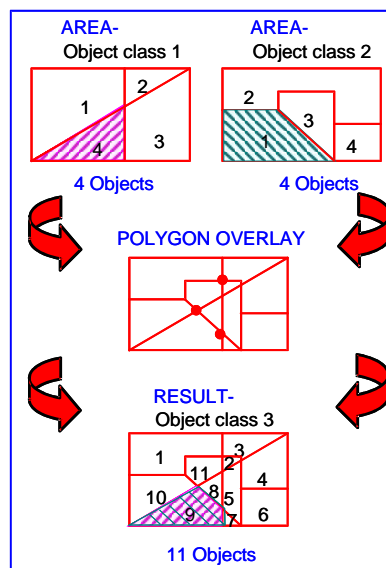


Figure 6: Polygon overlay (vector)

An overlay is a far easier computing task in the raster world than in the vector world because only the pixels involved in a particular case need to be interlinked (Figure 7)

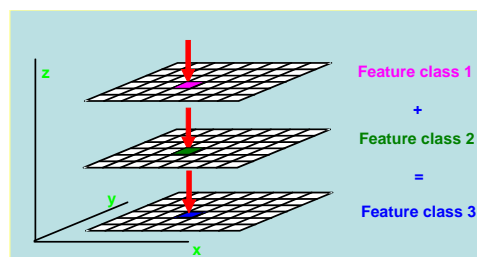


Figure 7: Polygon overlay (raster)

Questions such as “Which measurement points are situated within a community?” are answered by overlays of points and surfaces, which is equivalent to the above-mentioned point-in-polygon. The question “Which proportion of line y lies on lot x?” is answered, on the other hand, by an overlay of lines and polygons. “Which proportion of the lot is used for what purpose” is answered by an overlay of polygons and polygons. The geometrical overlay is accompanied by the transfer of attribute data from the source objects to the created objects. The reversal of a polygon overlay is called a **polygon dissolve**. Multiple attribute sets, which arise as the result of a polygon overlay, for example, are broken down again into their individual components.

**Triangulation** and **neighbourhood graphs** are analysis functions to move over from point features to triangles or polygonal features. In mathematics, and computational geometry, a **Delaunay triangulation** (Figure 8) for a set P of points in the plane is a triangulation DT(P) such that no point in P is inside the circumcircle of any triangle in DT(P). Delaunay triangulations maximise the minimum angle of all the angles of the triangles in the triangulation; they tend to avoid "sliver" triangles (Bill, 2016, en.wikipedia.org). Triangulation and neighbourhood polygons (called **Thiessen polygons** or **Voronoi diagrams**) are available for vector and raster data and allow the linking of geometrical methods with topological methods. Triangular irregular networks (TIN) form the base for terrain modelling.

Geometrical methods also deal with three dimensions (2.5D or 3D), such as **interpolation** (possibly with time 4D) to generate isolines, slope and gradient, exposition, aspect, hill-shading or line-of-sight. On the 2.5D surface we can compute surfaces, **network flows** and make **path analysis**. In 3D, for instance for geology, we can analyse the behaviour inside objects. We can create **buffers**, **cross-sections** or **profiles** in 3D or do **volume and deposit calculations**.

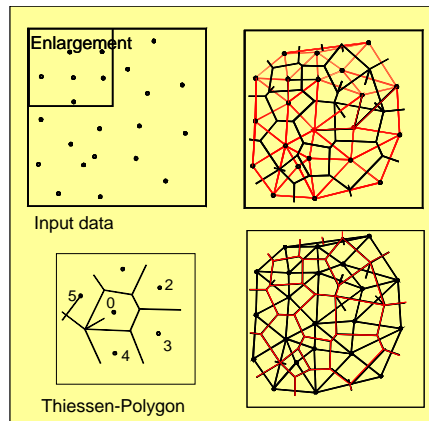


Figure 8: Delaunay triangulation and Thiessen diagrams

### 3.3.2 Topological methods

The mathematical background of topological methods is graph theory and the branch of topology dealing with adjacency and incidence. Topological algorithms and applications are to compute a best path, best site, or Travelling Salesman Problem. Graph theory offers a lot of shortest path algorithms such as the Floyd-Warshall-algorithm or the **Dijkstra-Algorithm** (Bill, 2016), which allow the solution of three different types of problems, for instance in network information systems (NIS).

**Shortest path:** In graph theory, the shortest path problem is the problem of finding a path between two nodes such that the sum of the weights of its constituent edges is minimised. An example, which we are all familiar nowadays using our navigation systems, is finding the quickest way to get from one location to another on a road map; in this case, the nodes represent locations and the edges represent segments of road and are weighted by the time or distance needed to travel that segment (en.wikipedia.org).

**Best site:** Determine the best site of a school, a new enterprise etc. in terms of reachability and commuter-belt.

**Travelling salesman:** The travelling salesman problem (TSP) in discrete or combinatorial optimisation tries to solve the following question: Given a number of cities and the costs of travelling from any city to any other city, what is the cheapest round-trip route that visits each city exactly once and then returns to the starting city?

**Example:** An example should illustrate the principal ideas of topology. Given five nodes A..E and their seven connections (the edges 1..7) shown here, what is the neighbourhood information and what shortest paths could be established?

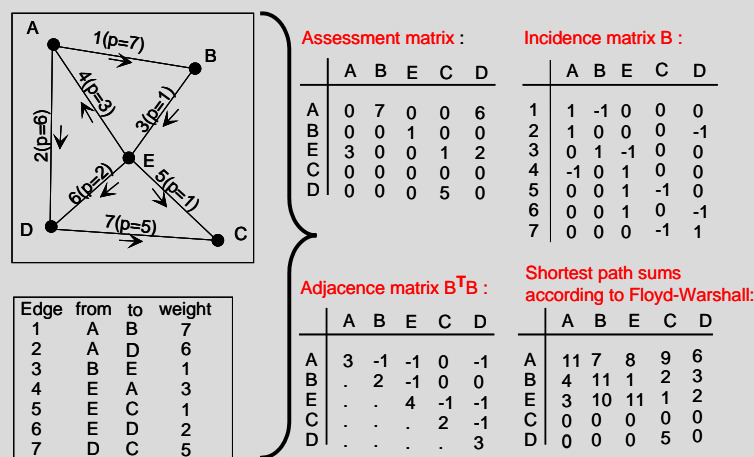


Figure 9: Shortest paths – incidence and adjacency

Searching for the nearest neighbour or shortest link between two locations, determining the course of a utility line between two locations, and carrying out complete topological and geometrical network analyses and simulations are all assignments that are generally based on topological information, They are often performed in both the transport logistics and energy supply sectors. They can be used just as readily to answer questions such as, “Which areas will suffer a temporary shortage in supply if there is a damaged pipe in the water system?”, “Which is the quickest way for a police car or fire brigade to reach the scene of an accident?” or “Which is the quickest and cheapest route for a parcel delivery vehicle to reach a customer?”.

### 3.3.3 Temporal methods

Temporal methods support the analysis of **landuse changes over time** by comparing time slots. In Europe the CORINE Land Cover (CLS) data set is available in the time slots 1990, 2000, 2006, 2012, and 2018 allowing to monitor the landuse changes. Similar to geometry and topology one can analyse the distance between (**temporal topological**) or the duration (**temporal geometrical**) of events. Objects may also be dynamic: Floating car data can be analysed to illustrate main **trajectories** driven in a city.

### 3.3.4 Set methods

Set methods are primarily related to attributes in GIS. The mathematical background is set theory, allowing mathematical operations between various sets of information. Under computational aspects these operations are combined under the term Boolean operations and embedded into standard programming languages.

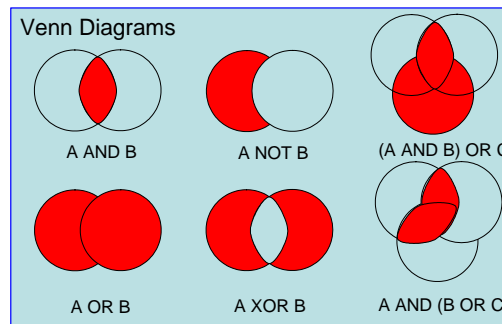


Figure 10: Boolean operators

**Boolean operations** (Figure 10) and **neighbourhood and connectivity analyses** combined with polygon overlays and abstractions support location planning. They serve to answer questions such as, “Where is the best location for a new supermarket, school or hospital?”, and they take account of demographic data such as age structure, income and consumer habits in addition to the topographical factors.

**Data retrieval** is the term used for the basic functions that can answer geometrical and topological queries addressed to the data – within a window, polygon or neighbourhood – as well as inquiries concerning specific attribute combinations. The data is combined selectively according to spatial (geometry and topology) and/or descriptive criteria (Figure 11) and the result is presented in graphic or some other form.



Figure 11: Selective queries

**Aggregation** is another very powerful analysis method. The term aggregation means combining data sets from different lower entities to higher entities. Very often data are collected at community level, but they are evaluated



at county or federal state level (Figure 12). Using a common key, for instance a unique community code, allows the aggregation of data from one level to the next simply by cutting some digits in the key and summarising the data sets only at that level. In contrast to the aggregation of spatial data, the spatial **disaggregation** assumes that available data of a given aggregation level can be distributed spatially within the boundaries of the area data by means of spatially differentiated parameters. The distribution is usually carried out via a weighted sum function. The basis for this is found in the statistical redistribution of socio-economic data. Well-known methods are, for example, the area interpolation method with or without additional data and the dasymetric mapping method.

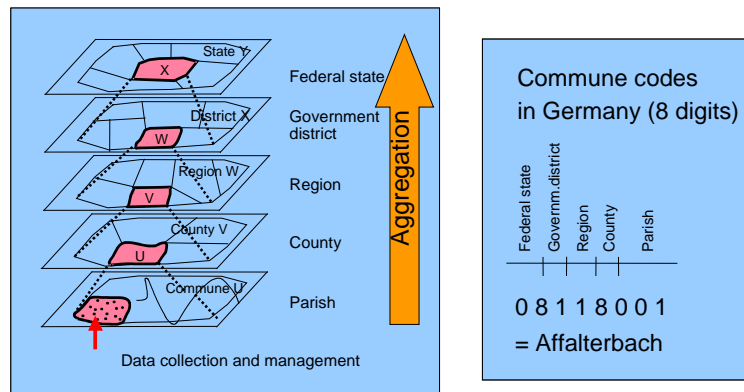


Figure 12: Aggregation

### 3.3.5 Statistical methods

Probability theory and stochastics offer powerful and sound theories for data analysis such as **descriptive statistics** or analytical statistics, either as univariate, bivariate or multivariate statistics. **Univariate statistics** offers characterisations of data sets computing the average or standard deviation and displaying histograms. **Bivariate statistics** analyses the correlation or covariance and the regression between two data series. **Multivariate statistics** has powerful methods such as **cluster analysis**, **factor analysis** and **multivariate regression**. These methods are outside of our focus here; they are described in textbooks on statistics.

In the context of GIS we are more interested in advanced **geostatistical methods** such as **interpolation** methods e.g. **variograms** and **Kriging**. Planar and spatial interpolations arise as basic functions of the digital terrain model, but they can also be applied to other data. In this case, data of one form are displayed in another form (clusters of points in contours or in neighbourhood graphs; clusters of polygons with uniform characteristics in their centroid points). Many interpolation functions exist for grid or triangular data sets such as TIN-Interpolation, interpolation with area summation, interpolation with minimum least squares methods, piece-wise linear polynomials, polynomial interpolation or Kriging. In this course we cannot discuss the mathematical foundations for these interpolation methods, but will simply use three diagrams to illustrate the power and effects of interpolation, comparing **inverse distance weighting** (Figure 13), **spline interpolation** (Figure 14) and **Kriging** (Figure 15).

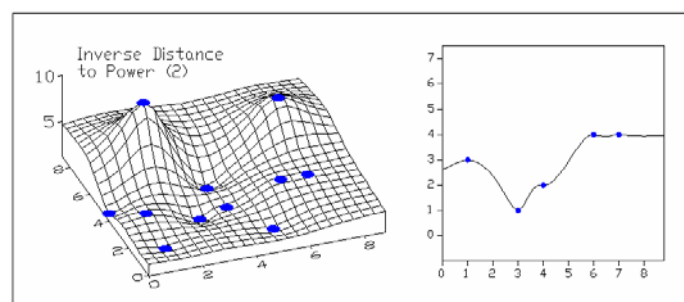


Figure 13: Interpolation with inverse distance weighting

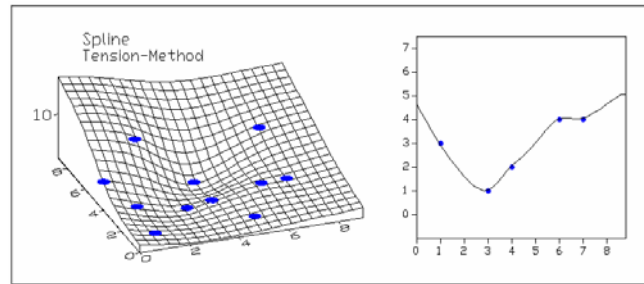


Figure 14: Interpolation with spline

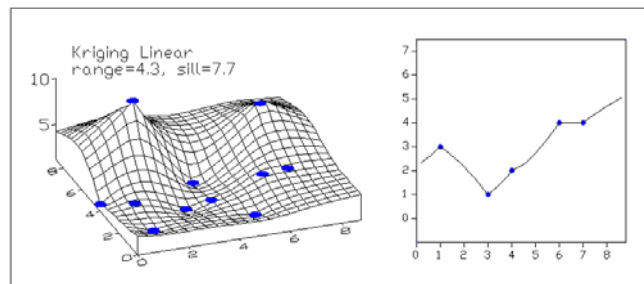


Figure 15: Interpolation with Kriging

### 3.3.6 Models and simulations

Many branch-specific and very complex models exist in GIS. They may be differentiated according to their geometric data type in point models (e.g. interpolation), line models (e.g. net flow calculations), or area models (e.g. dispersion). Geography offers many models, both on stochastic as well as deterministic assumptions of the phenomena to be studied. A very powerful tool to structure these complex models and processes into smaller units is the so-called map algebra, developed by C.D. Tomlin (1983), (1990) and made available with MAP, the **Map Analysis Package**, which divided a workflow into small parts that can be combined and offers a map algebra to process that workflow. Nowadays, similar methods are offered in GIS packages, for instance the Model Builder in ArcGIS.

Very often, related **simulations** and scenario calculations are performed outside the GIS, as does the generation of alternatives and assessment of environmental compatibility.

**Point models:** Functions for the acquisition and analysis of data that are dependent on the distribution of points and values over a plane are grouped together in the functions pool to form digital terrain models. They range from acquisition (primary data), meshing and digital terrain modelling (secondary data) to the creation of derived products such as contours and slope lines, 3D visualisations, visibility and illumination calculations, and much more besides (tertiary data). They find applications in all planning assignments, in the environment, forestry and land consolidation sectors, and in the development of infrastructures for commercial applications. A typical assignment for this function area is to answer questions such as, “Which is the most suitable point for locating a transmitter for the digital radio phone service if, in consideration of the given topography, we want to cover the largest area possible?”.

***Summary:** The data analysis component is the performance centre of a GIS. What is required is the most comprehensive range of functions possible and maximum flexibility and user convenience. The better the analysis tools provided by the GIS, the greater the benefits for the system user. Geometric operations are usually realised, Polygon overlay is therefore the basic function. Topologic operations are very important, but sometimes rather restricted. Functions for temporal analysis are getting more and more important. Set methods such as sort, search, query etc. are available. Simple descriptive statistics is realised, interpolations for DTM, geostatistics are rare or software and scripts outside the GIS. Models are usually externally realised for special applications.*

## 3.4 Data Presentation

The possibility of being able to work with a geo-information system on an interactive graphic level is a significant aspect of GIS utilisation. The most important basic functions that need to be provided to manipulate graphics on the screen of a computer are displaying, panning and zooming, each in conjunction with windowing. It must also be possible to create elaborate graphics and save them as a digital pattern collection for all the projects waiting to be processed. The interactive graphic design of such patterns is the main concern, and can differ very much from one output medium to another. Simpler patterns are preferred in interactive graphics in order to guarantee fast

image generation; map designers, on the other hand, want elaborate graphical symbolisation in order to make the maps they produce by digital means as identical as possible to those manufactured by analogue means.

There are many different graphic output or presentation forms (for further references see Part D in this book). In analogue form **maps**, detail- and **overview plans** and **sketches**, map **diagrams**, business graphics, **perspectives**, imagery, length-/transversal **profiles**, **reports**, **statistics** and **tables** may be used to illustrate phenomena modelled and results produced in a GIS. On the alphanumeric output side it should also be possible to generate reports and lists for presenting database extracts and thematic overviews. Ad-hoc inquiries for combing the data according to specific criteria using **SQL** are just as important as being able to process regularly recurring queries using macros, a facility that simplifies day-to-day work with the system. In electronic or digital form, the most important way of presenting information is in an interactive way on the display screen. Obviously modern computer graphics and visualisation offers nowadays many more possibilities to present information, e.g. multimedia combining reports, imagery and further media, as **fly-throughs**, **animation** or in **virtual- or augmented-reality environments**. It must also be possible to retrieve the data and results of a GIS for exchanging with other systems in order to enable integrated data processing within an organisation. This requires **data exchange interfaces** to the CAD, DTP, database, word processing and spreadsheet software. Data exchange with offices outside the organisation (from a subcontractor to a client and vice versa) is also performed these days by digital means, usually over the internet. In these cases, the analogue map serves simply as proof of completeness – as a rule, work continues immediately with the digital data.

***Summary:** The presentation of data is the geo-information system's "visiting card". Elaborate maps, tables, statistics and other output forms are the only way to achieve attractive and vivid results and ensure customers' satisfaction.*

## 4 Fields of application and markets

*Jack Dangermond, the CEO of the ESRI company, one of the leading GIS product development companies of the world, once stated in a presentation: "The application of GIS is only limited by the imagination of those who use it."*

Geo-information systems have been used in surveying offices and energy supply companies for more than four decades. These groups of users required mainly support in the acquisition, permanent management and presentation of data in map-like form. Their requirements as regards data analysis were very limited in the phase of data generation. Geo-information systems cover a far greater range of applications, however. It is worthwhile using a GIS wherever plans are drawn up and decisions made on the basis of maps, address pools or collections of administrative and other thematic data, whether in banks for the management and analysis of real estate stocks and customer data, in waste management companies for better utilisation of vehicle fleet capacities, in the regional analysis of health services or in smart farming, in the planning of locations for new factories or public institutions, in the analysis of election results, in official statistics and in many other sectors. Everyone nowadays uses for location-based decisions GIS with the smartphone. If we try to structure these application fields we may come to five application facets which we will now describe.

### 4.1 Land information systems (LIS)

Land information systems deal with the systematic capture and visualisation of all data that is related to a single piece of land. It shows all characteristic data of a region to enhance planning and development. This term was defined by the Fédération Internationale des Géomètres (FIG) at the beginning of the 1970s. Most of the developed countries did set up automated land registers combining the former map and register information they have on ownership and real estate. Important properties for LIS are a stringent permanent data management, a legally regulated data modelling, limited interactivity (simple static queries) and limited descriptive data (appropriate attributes for ownership and parcel size). Such systems were originally created in surveying and mapping disciplines and are applied in surveying, real estate management and local and state governmental topographic mapping. Very often they define the base maps (cadastral and topographic maps) for all other users.

### 4.2 Network information systems (NIS)

Network information systems (NIS) are instruments to capture, manage, analyse and present working materials related to a network topology in a uniform reference system. They are applied in utility companies' management of their facilities for electricity, gas, clear- and waste-water, telecommunication etc. Their characteristics are free data modelling, stringent permanent data management, high graphical interactivity and many descriptive and attribute data, for instance also on customers. Their origin dates back to the 1980s, when the first interactive graphic

system was introduced in the so-called AM/FM-sector (Automated Mapping/Facility Management). Municipalities manage their water supply and waste water networks with NIS, they create traffic plans and monitor the noise emissions in the city with the help of GIS.

### 4.3 Regional planning or statistical information systems (RIS)

Regional planning or statistical information systems (RIS) are instruments for decision support in spatial observations and tools for planning and development. They contain huge amount of data collected about population, economy and urban development, infrastructure, land use and resources for regional developments based on a common reference system (primary and secondary metrics). The characteristics of these application fields are free data modelling and management of permanent data and a high amount of descriptive data. Spatial analysis is a major component and a high degree of interactivity with flexible queries is required. Data is usually stored with a vector geometry model; hybrid systems e.g. for census raster are becoming more prevalent. Thematic cartography is the major output product of such systems. In municipalities, RIS are used to create housing and zoning plans, to draw landscape plans, to plan the economic development of the commune and to advertise for further establishing of enterprises. GIS in the form of RIS is used for administrative work, for instance for school registration, for elections etc. Communes and counties may set up local and regional web-based spatial data portals including all available spatial information and delivering better workflow-oriented services for their citizens and customers (eGovernment, including spatial components this is often called **geoGovernment**).

### 4.4 Environmental information systems (EIS)

Environmental information Systems (EIS) are advanced GIS for capture, storage, processing and presentation of environmental data, e.g. on hazards and pollutions, in space, time and context. These data are the basis for describing the status of the environment to decide on protective measures. Originally these systems were developed in environmental planning. Nowadays they also include sensor networks measuring environmental quality parameters in real-time. They need to support free data modelling, management of permanent data, high interactivity, simulation of environmental processes. Data is often 2D or 3D, 4D (including time) is not unusual in such systems, with a high amount of attribute data. The scale ranges from small to large, hybrid systems are dominating. GIS is only one software module in addition to environmental measuring and monitoring systems, report systems, early warning systems, etc. Environmental information systems are set up for the environmental media soil, water and air. They are used for nature and landscape protection planning and for urban climate monitoring.

### 4.5 Specialist information systems (SIS)

Beside the four application facets there are nowadays many more branches using GIS. Specialist Information Systems (SIS) or **branch-specific information systems** is an open special class of geo-information systems where we collect all applications that are not covered by the other classes. Compared to LIS, NIS, EIS, and RIS they have no generic characteristics. Examples are car navigation systems, telecommunication systems, hotel- and tourism-information systems, geo-marketing, military applications and many others. There are even cities managing their parking garage situation by means of GIS and send SMS to the drivers of the cars asking for a park lot.

In principal one can state that GIS is used wherever spatial data occur and spatial analysis is needed, or even more generally, that wherever a map is put on the wall to illustrate spatial phenomena a GIS is or can be used.

*Summary: In the past, the use of geo-information systems has tended to be restricted to local and specialised applications. They have served to perform special individual functions and have been less concerned with interacting with other functions and general IT. This is due partly to the fact that terms of reference and the possibilities for cooperation still have to be conclusively defined, which for many potential users has been reason enough not to move into geo-information systems. Another drawback has been that new users have had to do without the basic data, i.e. maps, around which their work has otherwise revolved. This situation has changed worldwide, with the development of nation-wide digital data bases – so-called **spatial data infrastructures (SDI)** and **open governmental data (OGD)** - by national offices responsible for surveying and mapping. In many application disciplines the changeover from analogue methods, i.e. work bases on paper maps, to computer-aided digital methods is realised. Administrative data is frequently also available open and in digital form.*

## 5 GIS products

There are thousands of GIS products on the market offering the necessary functionality for all four IMAP-components. As the availability of data increases, so the need for analysis functions grows. In addition, the need for data sharing and data dissemination demands the use of new technologies such as the internet. The user – and particularly the potential user – faces the enormous task of observing and understanding this market with all its

rapid changes and performance leaps. Numerous exhibitions provide an opportunity to view the products in brief demonstrations, and impressions can be intensified by visiting the suppliers. With the market being so transient it is advisable, however, to draw on the services of a competent and neutral consulting company. The costs of this service will be repaid very quickly when you decide on the right system.

## 5.1 Hardware, software and data

The quality and the power of a GIS for spatial data processing depend primarily on the hardware and software employed. Apart from the actual computer, the hardware consists of numerous peripheral units, which in the GIS sector extend far beyond what is understood under hardware in standard desktop computing. As regards software, the key elements consist of the computer's operating system, programming languages, graphics standards and data bases.

In fact, today **hardware** is not a topic anymore! GIS runs on high-performance hardware at relatively low cost, GIS follows the client-server architecture. GIS is available on stationary and mobile computers, new peripheral equipment is integrated particularly for data capture (GNSS, digital photogrammetry, mobile pen computers).

**Definition:** Software is understood to be all the immaterial elements of a data processing system, i.e. all the programs and data that can be used on the system. This includes the operating system, the programming language, the graphics standard, the data base language, etc.

Standards are now available for many elements of **software**. The demand for standards is based on the wish to ensure interoperability while reducing maintenance and frictional losses in heterogeneous computer environments. Basic software is actually not a topic anymore either! GIS runs on all types of operating systems, essentially Windows, Linux or Unix. GIS offers user-friendly interfaces with windows, icons, menus and pointers (WIMP). GIS supports common standards and is based on standard programming languages such as C, C++, or Java. GIS supports the database language **SQL**.

In the last decade there has been a trend towards open systems and standardisation following industry-wide implementation **standards** defined by the Open Geospatial Consortium<sup>1</sup> (OGC) and international **norms** approved by the International Organisation for Standardisation<sup>2</sup> (ISO). **Interoperability** is the key issue today.

The time of added value due to **data** started around the millenium! GIS as a hybrid system supports vector- and raster data. GIS runs on data bases for geometry and feature/object data. GIS offers relational, object-oriented and object-relational data models. Geoinformation's availability is permanently increasing, getting more and more open and becoming widely available via the internet.

## 5.2 GIS product categories

GIS has gone through some major development steps during the last fourty years. Starting as monolithic and very complex systems, which were used by specialists only, we now have very flexible and easy-to-use systems, even as components via the internet and on smartphones, which everyone is able to use (e.g. a routing system or a city guide on the internet). The costs for such systems and the time to learn to use them have decreased. The availability of data is increasing.

Nowadays GIS-products very often are distributed in a company according to the server-client concept. One or many large servers may be set up as database servers, map servers, GIS servers, which can be queried by desktop GIS or even mobile GIS components on tablet PCs or smartphones. Nevertheless, we still find large universal or desktop GIS, especially on those departments and offices, where data are captured day-by-day and individual information systems are maintained (LIS, NIS, EIS, RIS). But nowadays such components are typically part of a distributed GIS solution which is dedicated to the demands of the individual user. Much functionality is delivered via an intranet or even the internet (for further details see Part F).

## 5.3 Commercial versus open source GIS products

Today a user has the choice between many different software packages. In principal one may differentiate between commercial products such as ArcGIS (ESRI), MapInfo (Pitney Bowes), or GeoMedia (Intergraph) and Open source products such as QGIS, GrassGIS, gvSIG. Some short explanations on ArcGIS and QGIS as well established products worldwide.

---

<sup>1</sup> <http://www.opengeospatial.org>

<sup>2</sup> <http://www.iso.org>

### 5.3.1 ArcGIS

ESRI (Environmental System Research Institute<sup>3</sup>) is the world leader in GIS (geographic information system) modelling and mapping software and technology. ArcGIS is an integrated collection of GIS software products for building a complete GIS within organisations and enterprises (Gorr & Kurland, 2007). ArcGIS enables users to deploy GIS functionality wherever it is needed — in desktops, servers, or custom applications, over the internet, or in the field. ArcGIS and its modules support the whole IMAP processing chain in a GIS. ArcGIS is based on a modular component-based library of commonly used GIS software components called ArcObjects™. The family covers ArcGIS Desktop clients as a suite of GIS applications, ArcGIS Engine as a development component, Server GIS such as ArcGIS Server and mobile GIS such as Collector or Survey123. The new software is now called ArcGIS Pro.

ArcGIS Desktop covers different integrated applications:

- ArcCatalog (Geodata browser, Metadata editor)
- ArcToolbox (Data conversion, analysis and other functions)
- ArcMap (Data visualisation, map layout)
- Extensions such as ArcGlobe, ArcScene, or ModelBuilder for Geoprocessing.
- ArcGIS Desktop is available in three functional product levels:
- ArcView (ArcGIS ArcView for analysis) supporting data capture (Shape files, Personal Geodatabases), analysis, visualisation, mapping, output and reporting
- ArcEditor (ArcGIS ArcEditor for extended editing) offering data modelling, topology, multi-user feasibility, full functionality for geodatabases
- ArcInfo (ArcGIS ArcInfo for Geoprocessing) with further analytical functionality, geoprocessing, polygon processing, complete GIS (Input, Update, Query, Analyse, Mapping), for larger GIS infrastructures and GIS professionals.

### 5.3.2 QGIS

QGIS<sup>4</sup> (Quantum Geographical Information System) is one of the most common open source Desktop based GIS software to visualize, analyse, model and answer spatial issues. The first version was published in 2002. The actual Version is QGIS 3.8.0, which integrates, in contrast to QGIS 2, Python 3 in its processing interface. QGIS can be run on multiple operating system like Mac OS X, Windows, Ubuntu, Linux and Unix (download link<sup>5</sup>). QGIS integrates many extensions such as SAGA GIS and GRASS GIS. With QGIS everyone has a very powerful tool to treat various kind of spatial problems.

## 6 Trends and outlook

GIS theory, technology, and application domains continue to develop. What is now a subject of research in database, visualisation, data analytics, or internet technology and so on will soon find its way into GIS technology. This chapter briefly considers a number of current research and development themes, both for functionality (IMAP) and technology.

- **Input:** New data acquisition methods are becoming available. Digital photogrammetry has shown a rapid development in the last two decades: digital orthophotos are easy to produce and are becoming standard products for GIS usage. Digital terrain models derived by aerial laser scanning (ALS) or by digital 3D-line cameras are another example for the rapid developments in photogrammetry. Remote sensing satellites with high spatial resolution have been on the market since 2000 (for further details see Part B). Unmanned aerial systems (UAS) are approaching the GIS market offering very flexible and precise photogrammetric data acquisition. GNSS is more or less the standard for many data capture applications, GNSS and tacheometry as local positioning systems (LPS) are merging together. Terrestrial (TLS) and mobile laserscanning (MLS) are emerging technologies for precise 3D data capture for smaller areas bringing in the third dimension. Mobile GIS brings geoinformation outside in the field combining GNSS/LPS with smartphones or tablet PCs, possibly with an online wireless (WLAN, GSM, UMTS, ...) connection to the GIS database in the office (for further details see Part B).
- **Modelling and management:** Database technologies are an integral part of GIS. Additional to the classical relational model (see Part E), spatial extensions are available for classical database products (such as Oracle Spatial or PostgreSQL) or GIS products (such as ArcGIS Server by ESRI). New types of databases allow the handling of data streams from sensors or the management of real-time data. Data

<sup>3</sup> <http://www.esri.com>

<sup>4</sup> <https://qgis.org/>

<sup>5</sup> <https://qgis.org/en/site/forusers/download.html>

modelling makes increasing use of the standardized Unified Modeling Language (UML). The standardisation work on Geoinformation (ISO standard family 191xx) by ISO Technical Committee 211<sup>6</sup>, has established a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth. Specifications and implementations are defined by the Open Geospatial Consortium<sup>7</sup> to deliver interoperable spatial interface specifications that are openly available for global use. Spatial data infrastructures (SDI) and services are being created on different scales, either globally, nationally or for regional or municipality usage.

- **Analysis:** More and more complex analytical functionality is available. Processing chains may be modelled with map algebra techniques, for instance in the model builder of ArcGIS. Analysis functionality, until now the domain of desktop GIS, is going to be delivered as services (e.g. OGC Web processing service WPS) using internet technology and are embedded in standard workflows in enterprises. More and more data from different application domains are becoming publicly available, creating a need for semantic interoperability and associated ontological mapping algorithms and services.
- **Presentation:** Visualisation of spatial information is much more than simple maps. Challenges for visualisation are to be seen on the one hand when presenting spatial data on small devices such as smartphones or tablet PCs (mobile navigation system, location-based services). On the other hand, more and more visualisation techniques are available via internet web pages and services allowing much more interaction and new visualisation approaches such as multimedia, animation, augmented (AR) and virtual reality (VR) (For further details see Part D). Virtual Earth projects such as Google Earth are challenging the classical GIS products and vendors and making GIS a true mass-market phenomenon. This opening of a “consumer-GIS” market could produce startling changes in the global GIS market which until now has tended to be dominated by large institutions such as government agencies and their requirements.

GIS is still an emerging and ambitious technology developing very fast. In developing countries and emerging markets the GIS development may often be even much faster than in Europe and the United States. Many steps and hindrances may be missed out because of differing statutory regulations, better technology, easier access to data, better data quality etc. Nevertheless, it will still take considerable time to set up spatial data infrastructures and to link different applications and organisations together. Collaboration and well-defined strategies are necessary for successful implementations and setting up of larger, multidisciplinary GIS-projects.

## 7 Summary

A GIS links and relates multiple databases and maps together. It makes maps and databases interactive, combines data from various sources and turns data into information and information into knowledge by spatial analysis. A GIS visualises relations and patterns between features. It supports exchange and sharing of data/information and knowledge. GIS encourages cooperation and communication among different users. GIS is becoming more and more standardised and interoperable.

The power of a GIS comes from the ability to relate different information in a spatial context and to reach conclusions about this relationship.

GIS is not simply a computer system for making maps, although it can create maps at different scales, in different projections, and with different colours. **A GIS is an analytical tool.** The major advantage of a GIS is that it allows the identification of the spatial relationship between map features. A GIS does not store a map in any conventional sense, nor does it store a particular image or view of a geographic area. Instead, a GIS stores the data from which it is possible to draw a desired view to suit a particular purpose. GIS is more than just spreadsheets, statistics packages or drafting packages, CAD or database systems, GNSS etc. Instead it makes use of all of these packages and their functionalities to **analyse real world phenomena**. GIS is not a static map – neither in analogue (paper) nor in digital form. Maps are often a “by-product” of a GIS, because they are simply one way to visualise the results of spatial analysis (see Part D).

A GIS is able to:

- illustrate facts quickly and graphically,
- visualise comparisons between different options (highlight variations),
- support arguments and
- support interdisciplinary work.

A GIS is not able to:

- define the problems/tasks for the user,
- ensure suitability of the chosen data and process model,
- guarantee that the results make sense and
- to prevent the user from choosing a theoretically correct but too expensive/complicated option.

<sup>6</sup> <http://www.isotc211.org>

<sup>7</sup> <http://www.opengeospatial.org/>

## References

### Textbooks

- Bartelme, N. (2005): Geoinformatik. Modelle, Strukturen, Funktionen. Berlin-New York: Springer. 454 pages.
- Bernhardsen, T. (2002): Geographic information Systems. An Introduction, 3rd Edition. London: John Wiley & Sons. Inc, 448 pages.
- Bill, R. (2016): Grundlagen der Geo-Informationssysteme, Wichmann Verlag Offenbach, 6. edition, 866 pages.
- Burrough. P., McDonnell, R., Lloyd, C.D. (2015): Principles of Geographical Information Systems, Oxford University Press; 3rd edition, 432 pages.
- Chrisman, N. (2001): Exploring Geographic Information Systems. 2nd edition. London: John Wiley & Sons. Inc, 320 pages.
- Demers, M. (2011): Fundamentals of Geographical Information Systems, 4th Edition. London: John Wiley & Sons. Inc, 464 pages.
- De Smith, M.J., Goodchild, M.F., Longley, P.A. (2018): Geospatial Analysis - A comprehensive guide – 2018. <https://www.spatialanalysisonline.com/>
- Gorr, W.L., Kurland, K.S. (2007): GIS Tutorial. Updated for ArcGIS 9.2. Paperback: 374 pages, ESRI Press, USA.
- Tomlinson, R. (2007): Thinking about GIS. Geographic Information System Planning for Managers. Paperback: 254 pages. ESRI Press. USA.
- Zeil, P., Kienberger, S. (Eds., 2007): Geoinformation for Development. Bridging the divide through partnership. Wichmann Verlag, Heidelberg. 232 pages.

### Online resources

- <https://www.opengeoedu.de/> - German e-learning portal on GIS and open data. Last visited August 2019
- <http://www.gitta.info> – Swiss e- learning project in English. Last visited August 2019
- <https://www.esri.com/training/> - ESRI Training. Last visited August 2019
- <http://en.wikipedia.org/> - Wikipedia - the free encyclopedia. Last visited August 2019





## **Part B**

# **Georeferencing and GNSS**

Dipl.-Ing. (FH) M.Sc. Matthias Naumann, Dr.-Ing. Alexander Born and Prof. Dr.-Ing. Ralf Bill



# 1 Introduction

The term "geodesy" comes from Greek, can be translated as "division of the earth" and covers the entire field of surveying (Resnik/Bill 2018).

“The problem of geodesy is to determine the figure and the external gravity field of the Earth and other celestial bodies as functions of time; as well as, to determine the mean Earth ellipsoid from parameters observed on and exterior to the Earth’s surface” (Draheim 1971 and Fischer 1975 quoted in Torge 1991).

Reference systems are introduced to describe

- the motion of the Earth in space (celestial system), and
- the surface geometry and the gravity field of the Earth (terrestrial system).

“For global geodesy, the use of three-dimensional Cartesian coordinates in Euclidian space is adequate. In geodetic surveying, a reference surface is introduced in order to distinguish curvilinear surface coordinates and heights” (Torge, 1991).

## 1.1 Geodesy and its relation to other disciplines and sciences

For more than two hundred years, geodesy – strictly speaking, only one part of geodesy, i.e., positioning – was applied in mapping in the disguise known on this continent as “control surveying”. Positioning also finds applications in the realms of hydrography, boundary demarcation, engineering projects, urban management, environmental management, geography and planetology. At least one other part of geodesy, geo-kinematics, may also be applied in ecology.

Geodesy has a symbiotic relation with some other sciences. While geodesy supplies geometrical information about the Earth, the other geo-sciences supply physical knowledge needed in geodesy for modelling. Geophysics is the first to come to mind: the collaboration between geophysicists and geodesists is quite wide and covers many facets of both sciences. As a result, the boundary between the two sciences has become quite blurred, even in the minds of many geo-scientists. For example, to some the study of global gravity field fits better under geophysics than geodesy, while the study of local gravity fields may belong to the branch of geophysics known as exploration geophysics. Other sciences have similar, but somewhat weaker relations with geodesy: space science, astronomy (historical ties), oceanography, atmospheric sciences and geology.

As with all exact sciences, geodesy makes a heavy use of mathematics, physics, and of late, computer science. These form the theoretical foundations of geodetic science and thus play a somewhat different role with regard to geodesy. In Figure 16, we have attempted to display the three levels of relations.

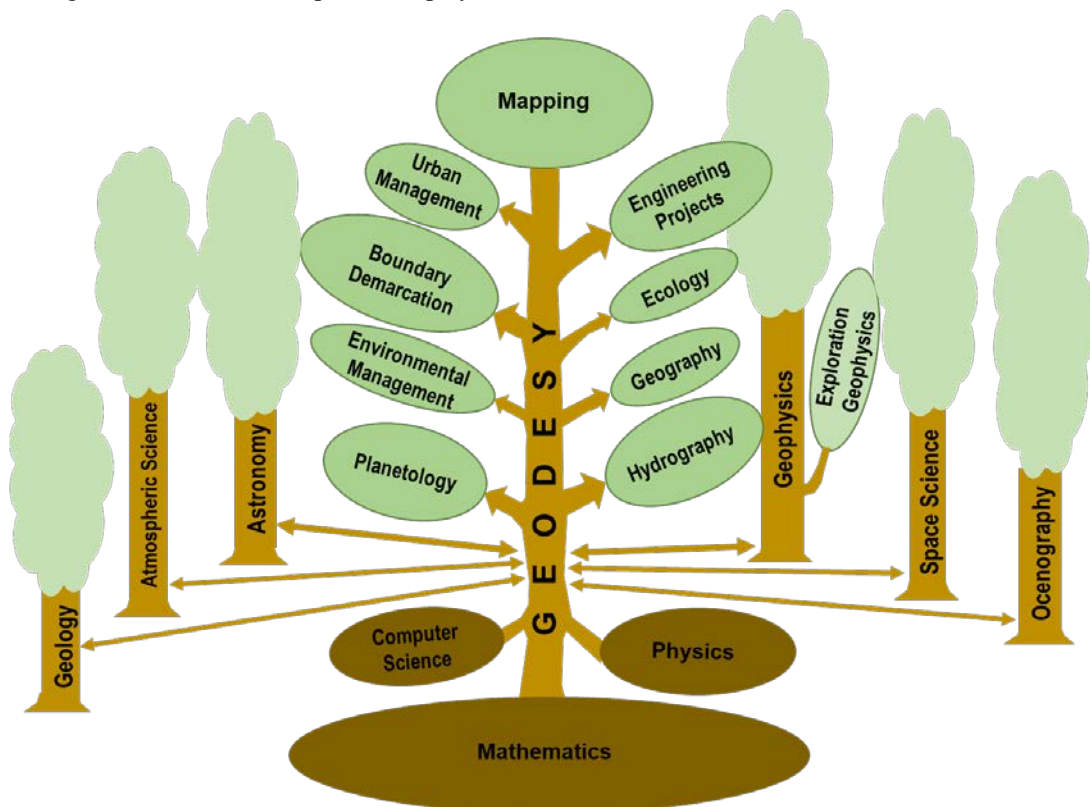


Figure 16: Geodesy and its relations to other disciplines (based on Vaniček and Krakiwsky, 1986)

## 1.2 The profession and practice of geodesy

Geodesy, as most other professions, spans activities ranging from purely theoretical to very applied. The global nature of geodesy dictates that theoretical work is done mostly at universities or government institutions. Few private institutes find it economically feasible to do geodetic research. On the other hand, it is quite usual to combine geodetic theory with practice within one establishment. Much of geodetic research is done under the disguise of space science, geophysics, oceanography, etc.

Of great importance to geodetic theory is international scientific communication. The international organisation looking after geodetic needs is the International Association of Geodesy (IAG), the first association of the more encompassing International Union of Geodesy and Geophysics (IUGG) which was set up later, in the first third of the 1900s. Since its inception, the IAG has been responsible for putting forward numerous important recommendations and proposals to its member countries. It also operates several international service providers such as the International Gravimetric Bureau (BGI), the International Earth Rotation Service (IERS), Bureau Internationale des Poids et Mesures – Time Section (BIPM), the International GPS Service (IGS), etc. The interested reader would be well advised to check the current services on the IAG web page<sup>8</sup>.

## 2 Physical, mathematical and geometrical fundamentals

### 2.1 Introduction

The figure of the Earth is approximated differently for the determination of heights and positions on or exterior to the Earth's surface. Because of its physical meaning, the geoid is well suited as a reference surface for heights. Terrestrial surveying instruments are aligned with spirit levels to the force of gravity, their measurements therefore refer automatically to the geoid.

The ellipsoid can be used as the basis for survey work, although one on which calculations are rather difficult. For general mapping applications, the Earth is depicted in two dimensions. The vertical component of a position is used for height measurements and the horizontal position is projected on a reference surface, which can be a plane, cone or a cylinder. This step, mapping geographic 2D-coordinates to a plane, is a special coordinate conversion (see section 3.4). There are numerous map projections, which are distinguished by the mathematical algorithms used to convert geographic coordinates into 2D projected Cartesian coordinates. In geodesy, surveying and navigation conformal projections, which use the surface of a cylinder, are frequently used. In a first step, the coordinates are projected onto the cylinder. In a second step, the surface of the cylinder is transformed to a plane by using specific projection methods and formulae. The disadvantage of almost any map is that it cannot exactly express reality and has certain distortion. Each map projection is always related to an underlying Earth model and its datum. If it is not related to the correct datum, large errors in horizontal position (hundreds of meters) may result. For this reason it is important to self-document the information about the spatial reference system of geodata and GIS resources by using associated metadata descriptions.

### 2.2 The shape of the Earth

Different approximations for the Earth's surface have been used throughout history. Before Newton began the spherical era of geodesy, the Earth had been regarded as being a **sphere** or a **plane**. The assumption that the Earth is a sphere is only possible for small-scale maps (smaller than 1:5,000,000). At this scale, the difference between a sphere and an ellipsoid is not detectable on a map (ESRI, 2004).

Although today it is well-known that the figure of the Earth is best described by the **geoid**, the Earth is represented by an ellipsoid to make mathematical calculations easier. Nevertheless, each measurement with an instrument aligned with a spirit level refers to the geoid.

The **ellipsoid** is only an approximation of the Earth's figure, which cannot be described without consideration of gravitation conditions. "What we call the surface of the Earth in the geometrical sense is nothing more than that surface which intersects everywhere the directions of gravity at right angles, and part of which coincides with the surface of the oceans" (C. F. Gauss 1828 quoted by Torge, 1991). „We may think of this surface as being extended under the continents and then identify it as the mathematical figure of the Earth“ (Helmert 1880 quoted by Torge, 1991). J. B. Listing designates this level surface or equipotential surface as the geoid.

---

<sup>8</sup> <http://www.gfy.ku.dk/~iag/>

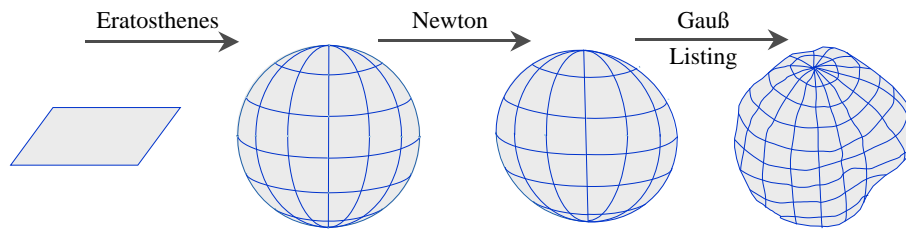


Figure 17: Geodetic reference figures (Resnik/Bill, 2018)

## 2.3 The geoid and vertical reference systems

### 2.3.1 The geoid

The geoid is used as a model for the description of the Earth's figure because it is better adapted to certain physical effects. "The potential of the Earth's gravity is alike at each place of the geoid's surface. The natural plumb-line direction and the geoid are located in each point perpendicularly to each other. Therefore the geoid can be determined by fairs of the force of gravity. Two arbitrary points on the geoid have the same weight potential and therefore the same dynamic height. The velocity of gravitation  $g$  is, however, not constant and decreases from the pole to the equator of 9.83 to 9.78m/s<sup>2</sup>." (Wikipedia, 2007)

With other words one can say that the geoid can be defined as the equipotential surface of the gravity field of the Earth and it is determined by the uneven distribution of the Earth's masses. In a good approximation, the geoid is represented by the mean sea level of the oceans and thereby, is visible in its form outside the land masses. Each resting water surface corresponds exactly to a section of such a reference surface.

The geoid is irregular, unlike the mean Earth ellipsoid (mathematical approximation of the Earth), but considerably smoother than Earth's physical surface. It is obvious that performing calculations on this "potato" is not simple. Therefore, the geoid is simplified into a more mathematical model, the ellipsoid. The mean Earth ellipsoid is the optimal ellipsoid approximating the geoid. Figure 18 shows the deviation of the geoid from the ellipsoid surface highly exaggerated.

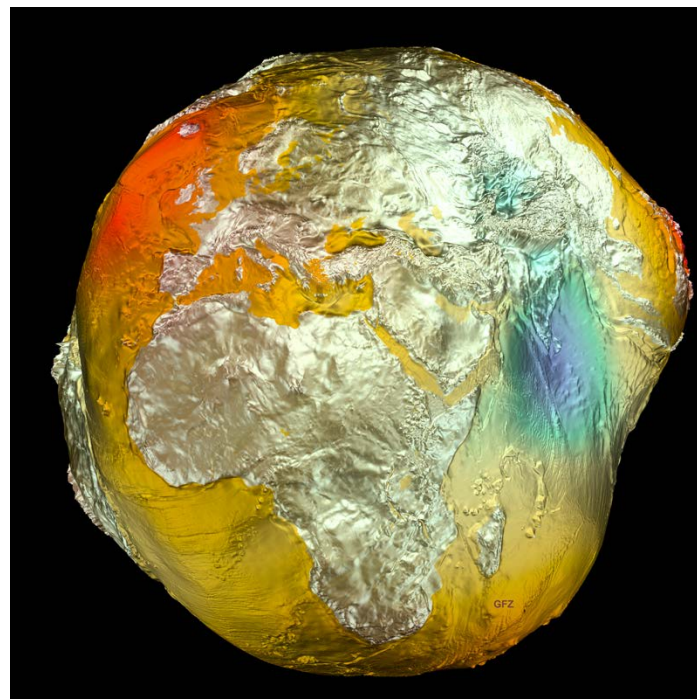


Figure 18: Distribution of the earth masses or the spatially non-uniform gravitational field of the earth in a highly exaggerated representation (German Research Centre for Geosciences (GFZ), 2011)

The figure of the Earth and the external gravity field are accordingly conceived as time dependent variables. This leads to the idea of "four-dimensional geodesy". The importance of the geoid has again increased with the establishment of three-dimensional continental and global systems, as well as with the requirements of marine geodesy (Torge, 1991).

### 2.3.2 Vertical reference systems for height measurements

The height is defined as a distance of a point from a chosen reference surface positive upward along a line perpendicular to that surface (ISO 19111:2019). Practical heights in geodesy are referred to the geoid (Figure 19). They are taking into account different assumptions about gravity between reference figure and surface point. There exists different possibilities to define the height:

- **Geopotential cote** defines the height of a point P as the negative potential difference to the geoid (height equals work).
- **Dynamic height** is computed based on the geopotential cote using a constant gravity value:  $H_{\text{dyn}} = \text{geopotential cote}/\text{constant gravity value}$ .
- **Orthometric height**  $H_{\text{orth}}$  describes the length of the curved perpendicular line from the terrain point to the geoid taking a mean gravity value along the perpendicular line. This was the former height system used in Germany named NN heights.
- **Normal height**  $H_{\text{norm}}$  measures the distance along the curved perpendicular line, of a point from the quasigeoid, a hypothesis-free defined reference surface, which corresponds to a "smoothed geoid" as an exact arithmetic surface. For the computation the mean normal gravity value along the perpendicular line is used. This height system is nowadays used in Germany and is called height above normal height zero (NHN).
- A geometric height definition using an ellipsoid as reference is the so-called **ellipsoidal height  $h$**  measuring the length of the ellipsoid normals between the point on the earth and the reference ellipsoid. These ellipsoidal heights are treated as related to a three-dimensional ellipsoidal coordinate reference system referenced to a geodetic datum (see section 2.6), but do not model physical behaviour on earth, i.e. the flow of water on the earth relief.

The quantity  $N$ , the height of the geoid above the reference ellipsoid – called **geoid undulation** – is usually called the geoidal height.

Tide gauges are used to define a geoid at a specific point. The corresponding vertical reference system is based on a vertical datum. As with horizontal datums (see section 2.6), there are many discrepancies among vertical datums. Precise geodetic levelling is used to establish a basic network of vertical control points based on the main vertical reference point (tide gauge). From the network points, the height of other positions in the survey can be determined by supplementary methods.

Since Real-Time Kinematic GPS (RTK-GPS) measurements with a high accuracy are possible, this method for the determination of the network of the vertical control points is rapidly gaining in popularity. As a requirement for this, accurate models of the geoid undulation  $N$  must have been determined by gravimetric measurements at many grid points. Global grid correction models for converting ellipsoidal heights to orthometric heights nowadays results from satellite missions, e.g. CHAMP, GRACE, GOCE and Swarm (<https://www.gfz-potsdam.de/sektion/geomagnetismus/infrastruktur/>). Their spatial resolution is constantly refined. For regional dimensions, gravimetric measurements are also taken from the aircraft.

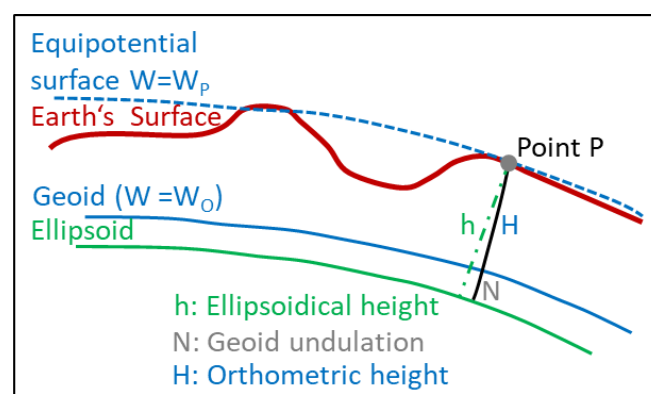


Figure 19: The relation between orthometric and ellipsoidal heights

**Summary:** A vertical datum describes the relation of gravity-related heights to the Earth. In most cases the vertical datum will be related to sea level. Vertical datums include “sounding datums” used for hydrographic purposes, in which case the heights may be negative heights or depths (ISO 19111:2007). Precise geodetic levelling is related to the geoid; on the other hand RTK-GNSS height measurements are based on ellipsoidal heights  $h$ . For conversions between both height systems one needs accurate models of the geoid undulation.

## 2.4 The reference ellipsoid

A reference ellipsoid is a mathematical-geometrical substitution for the surface of the geoid, which is too complex for computations. The ellipsoid, an ellipse rotated around its short axis, came into full development in the 19th century and today it is still the basis for geodesy, survey and cartographic work. There are several definitions of ellipsoids, they have been utilized in geodetic work and many are still in use. The older ellipsoids are named by their developers with the year of development. We distinguish between three kinds of ellipsoid definitions:

- **conventionally defined ellipsoids** (based on arc measurements),
- **regionally best-fitting ellipsoids** (based on astro-geodetic systems), and
- **global ellipsoids** (mean Earth surface approximations, geocentric equipotential ellipsoids).

In earlier geodetic surveys, the reference surface was a conventional ellipsoid computed from the adjustment of several arc measurements (e.g. Everest 1830, Bessel 1841). More recent geodetic surveys partly refer to a regionally best-fitting ellipsoid derived by the equations of the deflection of the vertical and geoid undulations for a large number of network points, using just one instead of four origin points. Computation for a geodetic datum from astro-geodetic deflections and geoid undulation can only be carried out on the continents. They provide ellipsoids which best fit the dimensions and positions of the geoid in their respective regions, the mean Earth ellipsoid cannot be determined in this way (e.g. Hayford 1924, Krassovsky 1940).

Since geocentric coordinates for a large number of points are available today – delivered with good values for the mean Earth ellipsoid – methods for determining astro-geodetic datums have lost their importance (Torge, 1991). Modern geodetic surveys are therefore based on the definition of global approximated and geocentered ellipsoids (e.g. GRS 80, WGS 84).

*Summary: There are different kinds of ellipsoid definitions (best-fit or Earth-centered). The ellipsoids differ in the parameters size (length of their axis), shape (the flattening), position (shift) and orientation (rotation) of the ellipsoid towards the Earth centre. These parameters are geometrical elements of the Geodetic Datum, which is - again - together with a coordinate system a component of the Coordinate Reference System (CRS).*

## 2.5 Datum

Each coordinate system is related to the real world by a datum (or reference frame), which is a set of parameters that may serve as a reference or basis for the calculation of other parameters. A datum is a set of parameters that defines the position of the origin, the scale, and the orientation of the axes of a coordinate system. It is the datum that makes the coordinate system and its coordinates unambiguous. A datum can be an engineering datum, a geodetic datum or a vertical datum.

- An **engineering datum** is a datum with a local reference used, for example, in surveying for a construction site or archaeological surveys.
- A **geodetic datum** describes the relationship of a coordinate system to the Earth and in most cases includes an ellipsoid definition (ISO 19111:2007). A geodetic datum defines the relationship of a geographic or geocentric coordinate system to the Earth. Attributes of a geodetic datum are the chosen model of the Earth (the ellipsoid) including details of name and defining parameter values (size, shape, position and orientation of the reference ellipsoid), together with the details of the zero- or prime meridian from which longitudes are measured (OGP, 2007c).
- A **vertical datum** describes the relationship of a gravity-related coordinate system (heights) to the Earth (ISO 19111:2007).

## 2.6 Coordinate systems

A coordinate system (CS) is a set of mathematical rules for specifying how coordinates are to be assigned to points. It includes the definition of the coordinate axes, the units which are used and the geometry of the axes. A coordinate system is unrelated to the Earth (IOGP, 2019). A CS is an abstract mathematical concept that is not tied to any physical or virtual object. A coordinate system is the set of coordinate system axes that spans the coordinate space (OGP, 2007c). The link between the Earth and a coordinate system is provided by the datum; both components define the Coordinate Reference System. Geodesy and cartography differentiate between ellipsoidal (or polar), cartesian and gravity-related coordinate systems.



## 3 Coordinate reference systems (CRS)

### 3.1 Introduction

The term coordinate system has historically been used to describe a CRS. A CRS is often incorrectly called a geodetic datum, but a geodetic datum is only one part of a CRS. It is important to note that a coordinate reference system always includes geodetic datum and coordinate system (Figure 20).

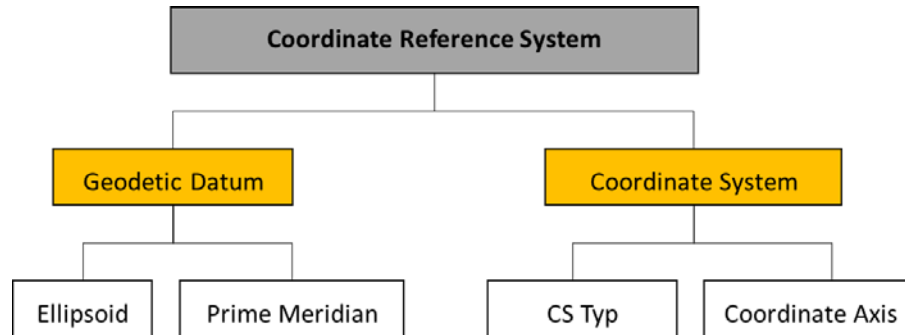


Figure 20: Components of a coordinate reference system (CRS)

A geodetic CRS uses a two- or three-dimensional coordinate reference system to describe spatial location over the whole Earth or substantial parts of it. The traditional separation of horizontal and vertical position has resulted in coordinate reference systems that are horizontal (2D) and vertical (1D) in nature, as opposed to truly three-dimensional. It is established practice to define a three-dimensional position by combining the horizontal coordinates of a point with a height or depth from a different coordinate reference system. In International Standards, this concept is defined as a compound coordinate reference system (2D+1D).

In addition, a geodetic coordinate reference system may also be three-dimensional (3D). When using GNSS to determine positions (see Chapter 5 “Surveying with GNSS”), three-dimensional geocentric-Cartesian coordinates are often used.

The International Standards ISO 19111:2019 (Geographic information -- Referencing by coordinates) differentiates types of CRS. In the context of spatial data processing, the following types are the most relevant and can be differentiated regarding their dimensions and their coordinate systems (CS) in:

- 3D: Geodetic CRS – geocentric Cartesian CS – geocentric X, geocentric Y, geocentric Z
- 3D: Geodetic CRS – geographic CS – geodetic latitude, geodetic longitude, ellipsoidal height
- 2D: Geodetic CRS – geographic CS – geodetic latitude, geodetic longitude
- 2D: Projected CRS – Cartesian CS – northing or southing, easting or westing
- 1D: Vertical CRS – vertical CS – depth or gravity-related height
- 3D: Compound CRS

A **one-dimensional CRS** indicates only the height, mostly gravity-related with respect to reference surface.

**Two-dimensional CRS** describes only the horizontal position, either as geographic ellipsoidal coordinates or as two-dimensional Cartesian coordinates projected on a plane surface.

A **three-dimensional CRS** describes the position in its three dimensions relative to the figure for the earth (mostly an ellipsoid, seldom a sphere), either as geocentric Cartesian coordinates X, Y, Z or as geodetic longitude, geodetic latitude and ellipsoidal height.

In the context of GIS, geographic CRSs with ellipsoidal coordinates, latitude and longitude are often used. In contrast, the explicit use of geocentric-Cartesian coordinates in GIS is not common. Although it is possible to create them with GNSS receivers, usually the geocentric-Cartesian X,Y,Z are converted beforehand into geographic coordinates, latitude and longitude.

**Map projections**, a special type of coordinate conversions, are used to derive projected coordinate reference systems (PCS) for a region. Map projections bring the curved earth's surface into the map plane. Latitude and longitude ellipsoidal coordinate values are converted into two-dimensional Cartesian coordinates. This creates three types of map distortions. Depending on the specific features supported by a map projection, the resulting map will be conformal to the angle, length-faithful or area-accurate. One or more projected coordinate reference system can be derived from a geographical coordinate system by use of one or more map projections (see section 3.3 and 3.4).

In contrast to the geodetic CRS, a three-dimensional Compound CRS connect a vertical to a projected CRS and separates the horizontal and vertical components. Both refer to different earth figures, the horizontal part

to the mathematical ellipsoid and the vertical part to the physical geoid. This combines the advantages of both systems.

One or more projected coordinate reference system can be derived from a geographical coordinate system by use of one or more map projections (see section 3.3).

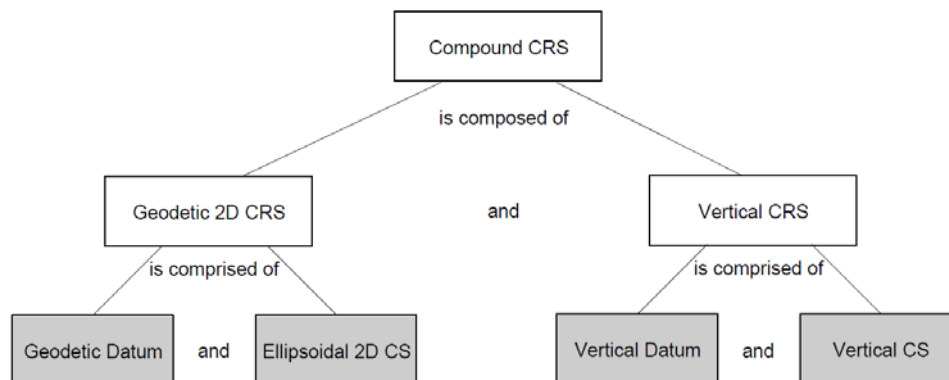


Figure 21: Conceptual model of spatial compound CRSs (OGC, 2010)

### 3.2 Geodetic coordinate reference systems (Geodetic CRS)

A geodetic CRS in the sense of the ISO 19111 is a 2D- or 3D-coordinate system which is related to the real world by a geodetic datum (ISO 19111:2007) (see Figure 22). The term geodetic includes the subtypes geographic 2D, geographic 3D and geocentric CRS (see section 3.1).

The Geodesic CRS with ellipsoidal-geographic coordinates is the base class in the GIS. Several abbreviations are used for this term: GEODCRS, GEODETICRS, GEOGS or GCS. The abbreviation according to international standard ISO 19162:2015 for text representations of CRS in the well-known text format version 2 is GEODCRS or GEODETICRS. From one geodCRS, many projected reference systems (projCRS or projectedCRS) can be derived by individual coordinate conversions, so-called map projections.

Two coordinate systems, illustrated in Figure 22, are used to define 3D positions in geodetic CRS:

- Geographic<sup>9</sup> CS (geodetic latitude  $\varphi$ , geodetic longitude  $\lambda$ , ellipsoidal height  $h$ ), and
- Geocentric CS (geodetic  $X$ , geodetic  $Y$ , geodetic  $Z$ ).

Converting coordinates between all subtypes of geodetic CRS based on the same geodetic datum gives a mathematically unique result (see section 4).

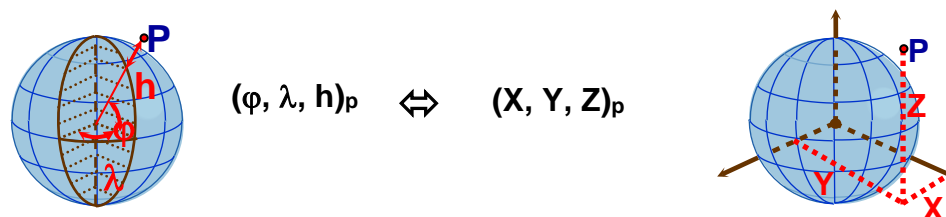


Figure 22: Ellipsoidal coordinate systems: geographical and Cartesian geocentric (Resnik & Bill, 2018)

For most practical applications, a geographic CRS is preferred because it facilitates a separation of horizontal position and height. In a geographic CRS, the Prime Meridian and the Equator are the reference planes used to define latitude and longitude:

- The **geodetic latitude** of a point is the angle from the equatorial plane to the vertical direction of a line normal to the reference ellipsoid.
- The **geodetic longitude** of a point is the angle between a reference plane and a plane passing through the point, both planes being perpendicular to the equatorial plane.
- The **ellipsoidal height** at a point is the distance from the reference ellipsoid to the point in a direction normal to the ellipsoid.

<sup>9</sup> ISO 19111 prefers the term *geodetic* (instead geographic) when the ellipsoid is related to the shape of the Earth. The ellipsoidal coordinates are called geodetic.

The geocentric-Cartesian CS is used to define 3D positions in a three-dimensional Cartesian coordinate system with respect to the centre of mass of the reference ellipsoid. The coordinates of a point are determined by X, Y and Z, usually using the unit meter. The coordinate axes are called geocentric X, geocentric Y and geocentric Z. The orientation of the axes is:

- The **Z-axis** points toward the North Pole.
- The **X-axis** is given by the intersection of the plane defined by the Prime Meridian and the equatorial plane.
- The **Y-axis** completes a right handed orthogonal system by a plane 90 degrees east of the X-axis and its intersection with the equator.

The **geographic 2D CRS** used in geodesy, surveying and middle- or large-scale cartography describes only position on the ellipsoid by using geodetic latitude and geodetic longitude. In small-scale cartography or for presentations in virtually globes often a sphere is used, in this case the coordinates are spherical latitude and spherical longitude. In geodesy applications the term ‘geodetic’ 2D CRS is often used instead of ‘geographic’ 2D CRS, this indicates that the reference surface is an ellipsoidal instead a spherical surface. Note that for some years the ISO has preferred the name geodetic.

Since EPSG dataset version 6.0 the CRS type ‘geographic’ has been replaced by the CRS type ‘geodetic’ which is subdivided into separate CRS types geographic 2D, geographic 3D and geocentric (for EPSG see section 3.5.1). However, in many GIS, the term geographic coordinate system (GCS) is still uses as a synonym for Geodetic CRS. This means that this geodetic CRS uses a geographic CS (with geodetic latitude and geodetic longitude) and refers to a specific geodetic datum.

It is important to describe coordinates with metadata about the geographic coordinate reference system used, because any given values of latitude and longitude can refer to any geodetic datum (see section 3.5). For example, about 5900 CRS and about 700 geodetic datums are registered in the EPSG Database (version 9.6) whose status is not ‘deprecated’ (see section 3.5).

**Example:** Since July 2000, “VN-2000” is the new national reference and coordinate system in Vietnam. It uses the global ellipsoid WGS 84 and the datum point N00. VN-2000 replaced the “Hanoi-72” system which in 1972 replaced the “Indian Datum 1960” (OGP, 2007a).

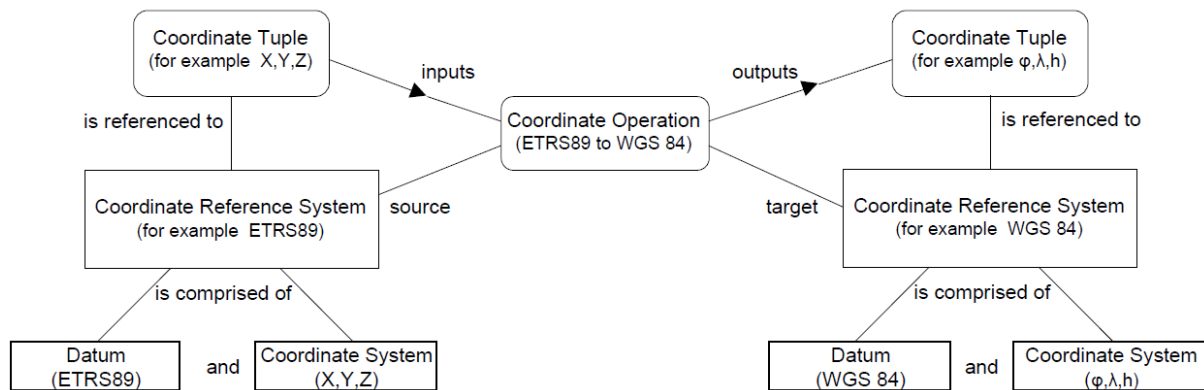


Figure 23: High level abstract model for spatial referencing by coordinates (source: IOGP 2019)

Converting coordinates between two geodetic CRS based on different geodetic datums (Figure 23) does not give a mathematically unique result. These operations are called transformation, to indicate the change between two datums (see section 4). Sometimes the transformation is called a datum transformation which is not strictly correct: not the datum, but the coordinates will be transformed.

### 3.3 Projected coordinate systems (PCS)

A projected CRS is the result of a map projection – a subtype of a coordinate operation method – to a geographic 2D-CRS (IOGP, 2019). A map projection uses mathematical formulas to relate ellipsoidal coordinates on the globe to flat, planar coordinates. This transformation creates a 2D coordinate system that considerably simplifies calculations between adjacent locations. Therefore, the term ‘map grid’ is also used. A projected CRS is described by two components, illustrated in Figure 24:

- the **projection method** used and its specific projection parameters, and
- the **geographic CRS** (in particular with its geodetic datum), which it is based on.

The most commonly used map projection methods preserve shape (conformal, see section 3.4). The coordinates in the projected CRS are sometimes designated as grid coordinates and called e.g. Easting and Northing.

Each projected CRS is based on a geographic CRS (and therefore on a datum). Vice-versa, a geographic CRS can be associated with many projected CRS. One map projection may be applied independently to many geographic CRSs (IOGP, 2012).

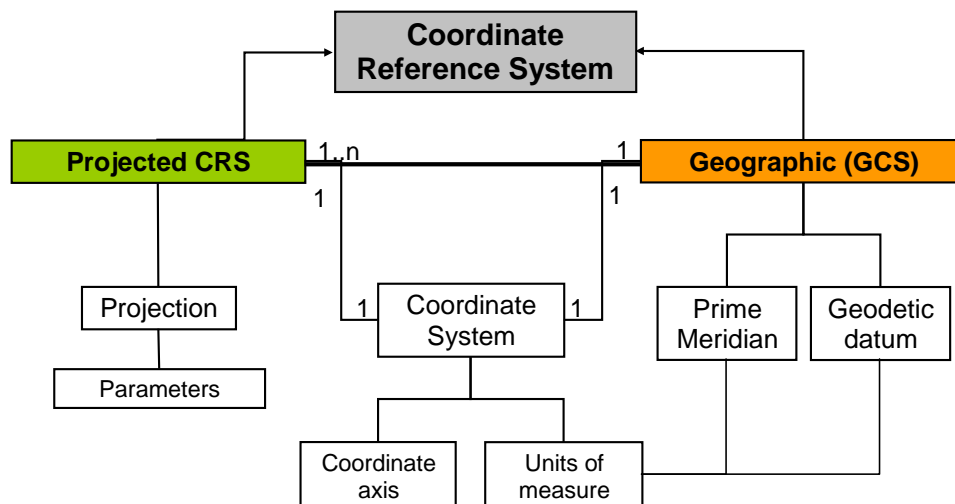


Figure 24: Relationship between GCS and PCS

The Geographic CRS “GCS\_Hanoi 1972” is the base for the projected CRS “Hanoi 1972 / Gauss-Kruger zone 18” and “Hanoi 1972 / Gauss-Kruger 106NE” (Figure 25). Both projected CRS use the Gauss-Kruger projection but different central meridians (105° and 106°) as an essential part of the projection parameters.

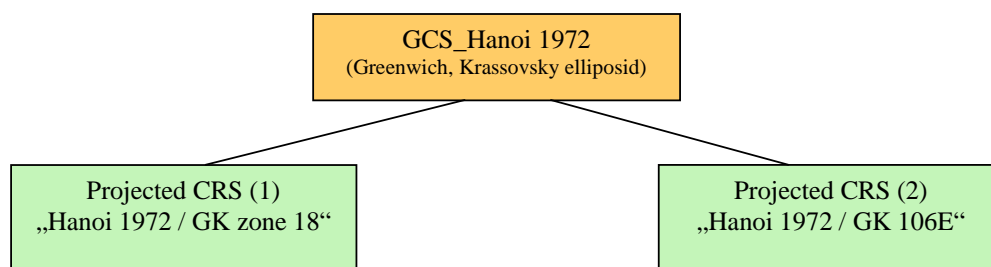


Figure 25: Geographic CRS and derived projected CRS (example Hanoi 1972)

### 3.4 Map projection

A variety of different map projections have been developed. The reason is that it is impossible to translate a curved surface into the plane without distortion. Only two of the features conformable, true to length, or true to area can be preserved. Small scale maps use a spherical earth and the projection formulas are simpler as for the ellipsoid. Projections of the sphere are only suitable for illustrative maps at scale of 1:1 million or less where precise positional definition is not critical. There are a large number of map projection methods, which may be employed for atlas maps, wall maps of the world or continental areas (IOGP, 2019). For medium and large scale sheet map projections, or maps and coordinates held digitally to a high accuracy, coordinate reference systems based on an ellipsoid and its derived map projections are necessary. For large scale topographic maps orthomorphic or conformal map projections are used (IOGP, 2019).

The geodetic coordinates (latitude, longitude) refer to the curved surface of the ellipsoid. For mapping purposes, the necessary projection into the plane can be performed by different projection types. A map projection is a type of a coordinate conversion (see section 4) which uses a mathematical function to convert ellipsoidal coordinates (excluding the heights) into two-dimensional Cartesian coordinates (plane coordinates), or vice-versa (ISO 19111:2019). The derived projected CRS is always related to its base geodetic (geographic) CRS and inherits attributes of the base geodetic (geographic) CRS, for example geodetic datum and ellipsoid (IOGP, 2012).

An intermediate step is to project the reference surface (or a part of this) on to a mapping surface, which can be a plane, cylinder or cone (Figure 26).

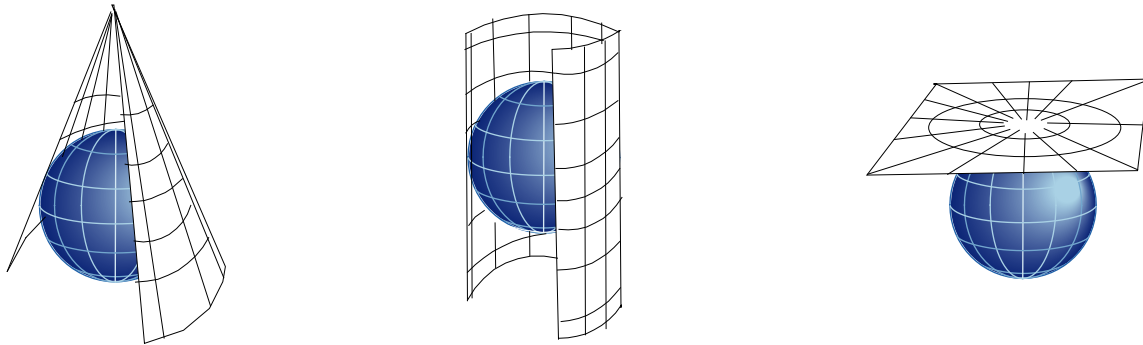


Figure 26: The three classes of map projections: conical, cylindrical and azimuthal. The projection planes are respectively a cone, cylinder and plane (Resnik/Bill, 2018).

Depending on the orientation of the plane, cone or cylinder relative to the Earth one speaks of normal, transversal and oblique orientation. Normal is by definition when the longitudinal axis of the cylinder coincides with the direction from north to south pole. Transversal is rotated by 90 degrees, so that the longitudinal axis is in the equatorial plane. Oblique denotes any inclination of the axis of the cylinder or of the cone or, in the case of the azimuthal projection, the arbitrary position of the projection plane.

An azimuthal map projection will have an associated reference point called the projection origin. Conical or cylindrical projections can have one or two touch meridians, depending on whether the Earth touches or partially penetrates the chosen body. Each map projection uses specific formulae and a set of parameters specific to that kind of map projection.

The disadvantage of almost any map is that it cannot exactly express reality and has certain distortions. Mapping of an ellipsoid into a plane can be conformal (or area-preserving), but length-preserving is not possible. In order to limit distortions within a certain range, only regions of limited extent around the point of origin or the central meridian are mapped.

For further readings see the textbook “ArcGIS 9: Understanding map projections” (ESRI, 2004) or the “Guidance Note Number 7, part 2” (IOGP, 2019).

It is important to note that any map projection, including UTM, may be applied with any geodetic datum. Therefore, projected CRS must be properly identified to avoid ambiguity (OGP, 2007b). For the description of coordinates belonging to a projected CRS, the coordinate description including geodetic datum, projection method and their parameters should be mandatory (see section 3.5).

The EPSG Dataset (see section 3.5) and this supporting conversion documentation usually considers only map projection for the ellipsoid. The following list of named map projection methods are those which are most frequently encountered for medium and large scale mapping. They are grouped according to their possession of similar construction properties and except where indicated all are conformal (IOGP, 2019).

- Lambert Conical Conformal (Conical) : with one standard parallel or two standard parallels,
- Mercator (Cylindrical): with one standard parallel or two standard parallels,
- Cassini-Soldner (Transverse Cylindrical, but not conformal)
- Transverse Mercator Group (Transverse Cylindrical)
  - Transverse Mercator (including south oriented version)
  - Universal Transverse Mercator
  - Gauss-Kruger
  - Gauss-Boaga
- Oblique Mercator Group (Oblique Cylindrical)
  - Hotine Oblique Mercator
  - Laborde Oblique Mercator
- Stereographic (Azimuthal)
  - Polar
  - Oblique and equatorial

### 3.4.1 Transverse Mercator (Gauss-Kruger) and Universal Transverse Mercator (UTM)

The Transverse Mercator projection in its various forms is the most widely used projected coordinate system for world topographical and offshore mapping (OGP 2007d). The surface of the cylinder is transformed to a plane by using specific projection formulae.

In the conventional **Transverse Mercator (TM)** projection (Figure 27) the standard meridian (or “central meridian”) is mapped without distortion as it is the line of tangency of the spherical approximation of the ellipsoid with the cylinder. The central meridian is the y-axis (north direction) of the projection, while the x-axis is the mapping of the equator. As a point moves away from the central meridian the distortions become larger as the scale factor increases.

The Gauss-Kruger projection is a special type of the Transverse Mercator and is used for instance in the former USSR, Germany, Yugoslavia, South America, China and Vietnam. There are variations in the width of the longitudinal zones for the Transverse Mercator projections used in different territories (usually 3° or 6°). The values of False Easting are various, but often 500,000m prefixed by zone number (IOGP, 2019) is chosen. The false northing is various.

“A UTM or other Transverse Mercator projection zone will normally extend only 2 or 3 degrees from the central meridian. Beyond this area another zone of the projection, with a new origin and central meridian, needs to be used or created.” (IOGP, 2019).

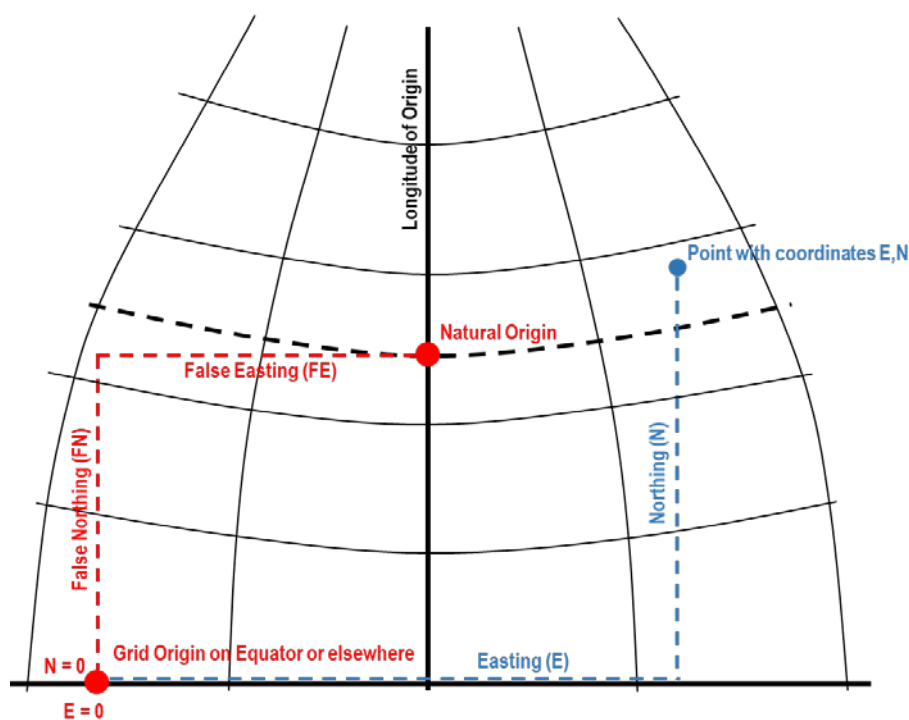


Figure 27: Transverse Mercator Projection parameters in Northern Hemisphere (based on IOGP, 2019)

Thus for projections in the Transverse Mercator group, the parameters which are required to completely and unambiguously define the projection method are (IOGP, 2019):

- Latitude of natural origin ( $\varphi_0$ )
- Longitude of natural origin (the central meridian) ( $\lambda_0$ )
- Scale factor at natural origin (on the central meridian) ( $k_0$ )
- False easting (FE)
- False northing (FN)

A further modification of the TM is the **Universal Transverse Mercator (UTM)** (Figure 28). UTM is an important projection in surveying and cartographic applications because it is an internationally-adopted map projection system. Firstly, the ellipsoid is partitioned into 60 zones, each 6° longitude in width. Secondly, the scale factor at the central meridian is 0.9996 to reduce the large distortions in the fringes of the zone (see Table 3 for all parameters). The UTM projection is defined between 80°S and 84°N. Outside these limits, the Universal Polar Stereographic (UPS) projection is used.

In contrast to areas north of the equator, for mapping southern hemisphere areas the equator origin is given a false northing of 10,000,000 m, thus ensuring that no point in the southern hemisphere will take a negative northing coordinate (IOGP, 2019). Figure 28 illustrates the UTM arrangements.

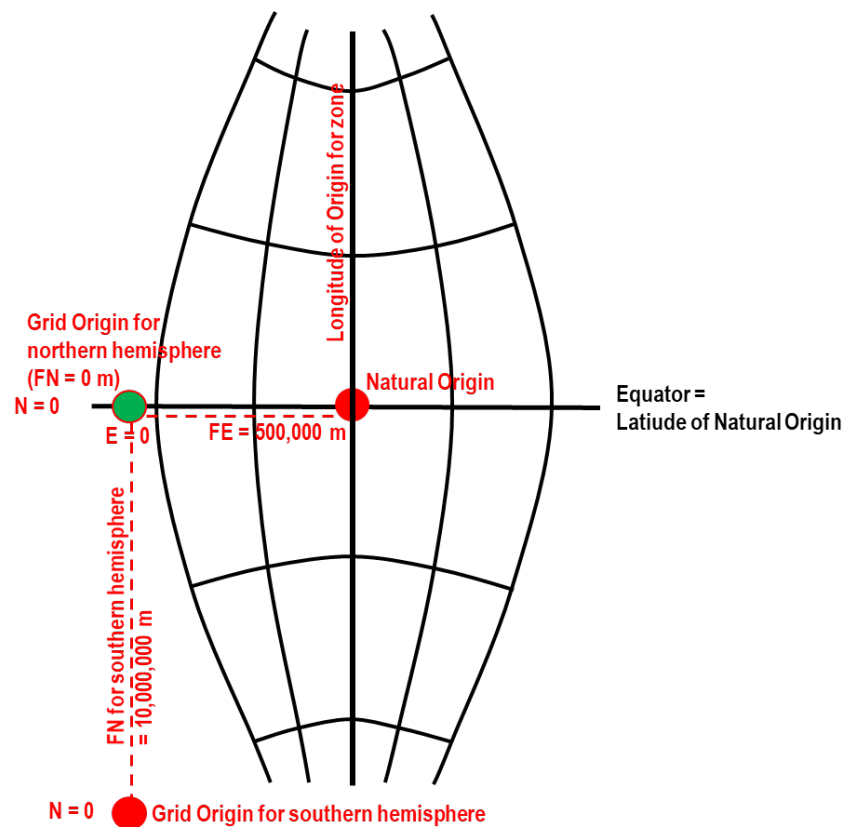


Figure 28: UTM Projection parameters in Northern and Southern Hemispheres (based on IOGP, 2019)

- Name	- Area	- Central meridian	- Latitude of natural origin	- Central meridian scale factor	- Zone width	- False Easting	- False Northing
- UTM North hemisphere	- World wide equator to 84° N	- 6° intervals E & W of 3° E & W	- Always 0°	- Always -0,9996	- Always - 6°	- 500000m	- 0m
- UTM South hemisphere	- World wide equator to 80° S	- 6° intervals E & W of 3° E & W	- Always 0°	- Always -0,9996	- Always - 6°	- 500000m	- 10000000m

Table 3: Parameters of the Universal Transverse Mercator (UTM) projection (IOGP 2007b)

### 3.5 Metadata descriptions of the spatial reference system

Each GCS and PCS is always related to an underlying Earth model and its geodetic datum. If it is not referenced to the correct geodetic datum or if it is used with a different prime meridian, it may result in large errors in horizontal position (hundreds of meters). The same will be the case with the parameters of the projection. Therefore it is imperative to describe data using meaningful digital information for their spatial reference and projection parameters.

#### 3.5.1 EPSG Geodetic Parameter Dataset (EPSG Dataset)

The most important register in this area is the so-called EPSG database (<http://www.epsg.org/>), abbreviated to the *EPSG dataset*, which is maintained by the Geodesy Subcommittee of the Geomatics Committee of the “International Association of Oil & Gas Producers” (IOGP). The IOGP’s EPSG Geodetic Parameter Dataset is a

structured collection of definitions of coordinate reference systems and coordinate transformations between different coordinate reference systems (<http://www.epsg.org/>) and it is distributed in three ways:

- the EPSG Geodetic Parameter Registry (abbreviated as *EPSG Registry*), a web-based delivery platform in which the data is held in GML using the CRS entities described in ISO 19136.
- the EPSG Geodetic Parameter Database (abbreviated as *EPSG Database*), a relational database structure where the entities which form the components of CRSs and coordinate operations are in separate tables, distributed as an MS Access database;
- in a relational data model as SQL scripts which enable a user to create an Oracle, MySQL, PostgreSQL or other relational database and populate that database with the EPSG Dataset.

In addition to the EPSG Dataset, the IOGP provides the Geomatics Guidance Note 7. This multi-part document provides detailed information about the EPSG Dataset and its content, maintenance and terms of use, as well as explanations of the formulas necessary for executing coordinate operations supported in the EPSG Dataset.

For example, the EPSG codes for the CRSs based on the World Geodetic System 1984 (WGS 84) datum are 4326 (geographic 2D), 4979 (geographic 3D) and 4978 (geocentric). In ESRI's GIS Software ArcGIS the EPSG code of a specific CRS is also called as well-known IDs (WKID) or factory code or authority code.

### 3.5.2 Well-known text compliant with ISO 19162

The former version of well-known text (WKT-1) standard based on ISO 19125:2004 described spatial reference system as string representation. In this older version, the coordinate reference systems are inconsistent. In addition, the different vendors implemented in different ways, so this leads to incompatibilities in data exchange.

Version 2 of the Well-known text (WKT-2) representation of coordinate reference system based on ISO 19162:2015. "ISO 19162:2015 defines the structure and content of a text string implementation of the abstract model for coordinate reference systems described in ISO 19111:2007 and ISO 19111-2:2009. The string defines frequently needed types of coordinate reference systems and coordinate operations in a self-contained form that is easily readable by machines and by humans. Because it omits metadata about the source of the data and may omit metadata about the applicability of the information, the WKT string is not suitable for the storage of definitions of coordinate reference systems or coordinate operations" (<https://www.iso.org/standard/63094.html>).

Table 4 shows a comparison of the same projected CRS coded in version 1 and version 2 of well-known text. The names in WKT version 2 are from EPSG Geodetic Parameter Registry. The parameters for the coordinate operations are now contained within. The PROJECTION keyword is deprecated in favour of METHOD.

Sample projected CRS coded in WKT-1:	Sample projected CRS coded in WKT-2:
<pre>PROJCS["WGS84 / Pseudo-Mercator", GEOGCS["WGS84", DATUM["World Geodetic System 1984", SPHEROID["WGS84", 6378137.0, 298.257223563]], PRIMEM["Greenwich", 0.0], UNIT["Degree", 0.0174532925199433]], PROJECTION["Popular Visualization Pseudo Mercator"], PARAMETER["False Easting", 0.0], PARAMETER["False northing", 0.0], PARAMETER["Longitude of natural origin", 0.0], PARAMETER["Latitude of natural origin", 0.0], UNIT["meter", 1.0]]</pre>	<pre>PROJCRS["WGS 84 / Pseudo-Mercator", BASEGEODCRS["WGS 84", DATUM["World Geodetic System 1984", ELLIPSOID["WGS 84",6378137,298.257223563,LENGTHUNIT["metre",1.0]]], CONVERSION["Popular Visualisation Pseudo-Mercator", METHOD["Popular Visualisation Pseudo Mercator",ID["EPSG",1024]], PARAMETER["Latitude of natural origin",0,ANGLEUNIT["degree",0.01745329252]], PARAMETER["Longitude of natural origin",0,ANGLEUNIT["degree",0.01745329252]], PARAMETER["False easting",0,LENGTHUNIT["metre",1.0]], PARAMETER["False northing",0,LENGTHUNIT["metre",1.0]], CS[cartesian,2], AXIS["easting (X)",east,ORDER[1]], AXIS["northing (Y)",north,ORDER[2]], LENGTHUNIT["metre",1.0], ID["EPSG",3857]]</pre>

Table 4: Comparison of the same projected CRS coded in WKT-1 and WKT-2. (<http://www.epsg-registry.org>)

### 3.5.3 Others

There are various standards of documentation of the parameters of the spatial reference system in digital form together with the geodata, e.g.



- The **ESRI projection file** is a proprietary metadata format for the documentation of the Coordinate Reference Systems in ESRI GIS software. In case of the vector based ESRI shapefile format, the projection file is an additional plain text file associated with the shapefile and provides detailed information about the coordinate reference system and the map projection. The keywords used in ESRI projection files corresponds to the version 1 of well-known text format (WKT1), but at least the file contain the specific EPSG code.
- The **GeoTIFF** format is a quasi-standard for the georeferencing as well as the description of TIFF images with specific metadata (coordinate reference system, projection parameters) in the header section of these images. It is fully compliant with the widespread raster file format TIFF (specification version 6.0) and allows georeferencing information to be embedded within a TIFF file. All GeoTIFF specific information is encoded in several additional reserved TIFF tags, so a GeoTIFF file seems to be a TIFF 6.0 file. GeoTIFF enabled software can read the specific metadata that describes the georeferencing information, catering to geographic as well as projected coordinate systems needs (Mahammad et al., 2009).

Furthermore, there exist various formats for describing geographic information and services. They provide information about the identification, extent, quality, spatial and temporal schema, spatial reference, and distribution of digital geographic data, e.g.

- The “International Organization for Standardization” (ISO) technical committee TC211 develops the ISO standard 19115 “Geographic Information-Metadata”. “ISO standard 19139 ‘Geographic Information-Metadata—Implementation Specification’, provides an XML schema that says how ISO 19115:2007 metadata should be stored in XML format. Many countries, regions, and communities are adopting profiles of ISO 19115:2007 or ISO 19139 as their national standard” (ESRI, 2007).
- The Federal Geographic Data Committee’s (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) provides a complete description of a data source for the United States. “Because it is quite detailed, other states and regions have created their own metadata standards to try to simplify the information that should be recorded” (ESRI, 2007).

*Summary: Metadata should be stored as structured data (e.g. XML), either in a file alongside the item, or within its geodatabase. The structure and availability of the metadata elements differs in the various metadata formats. Applicability and appropriate use [of geodetic codes and parameters] are of great concern, as geographic information systems become more widely available to non-experts in cartography and geodesy (ISO 19127:2005). The EPSG codes are increasingly accepted by software vendors. Due to the short key number, the complex relationships of the descriptions of the CRS can be easily exchanged..*

## 4 Coordinate operations

It is a frequent requirement to convert coordinates derived in one geographic coordinate system to values expressed in another. For example, nowadays topographic survey points are most conveniently positioned using GPS in the WGS 84 geographic coordinate reference system, whereas coordinates may be required in the national geodetic reference system. It might therefore be necessary to transform the observed WGS 84 data to the national geodetic reference system in order to avoid discrepancies caused by the change of geodetic datum.

### 4.1 Definition: conversion versus transformation

The coordinate reference system (CRS) is an aggregated class with the component classes geodetic datum and coordinate system. A coordinate operation is a change of coordinates, from one coordinate reference system to another. Coordinate transformations and coordinate conversions are subtypes of coordinate operation (ISO 19111:2019).

A coordinate **conversion** is a change of coordinates from one CRS to another based on a one-to-one relationship (ISO19111:2019). It only changes the coordinates from one CS to another and does not change the geodetic datum. A map projection is the most frequently encountered type of coordinate conversion. A map projection effectively defines the projected CRS and inherits attributes of the base geodetic (geographic) CRS, for example datum and ellipsoid. The derived projected CRS is always related to its base geodetic (geographic) CRS (IOGP, 2012).



Figure 29: Process flow diagram for a coordinate conversion (based on Ihde et al., 2012)

A coordinate **transformation** is a change of coordinates from one CRS to another CRS based on a different datum through a one-to-one relationship (ISO19111:2019). Transformation parameters are empirically determined and thus subject to measurement errors. It is important to distinguish, unlike projections where creates derived projected CRSs, transformations always operate on coordinates referenced to two already-defined CRSs. The corollary is that transformations have no place in the definition of a CRS (IOGP, 2012).



Figure 30: Process flow diagram for a coordinate conversion (based on Ihde et al., 2012)

A **concatenated operation** is a series of transformations and/or conversions executed in sequence (IOGP, 2012). Figure 31 shows various concatenated operations. At first two steps of conversions, then a transformation between both geodetic CRSs and again two conversions in reverse order.

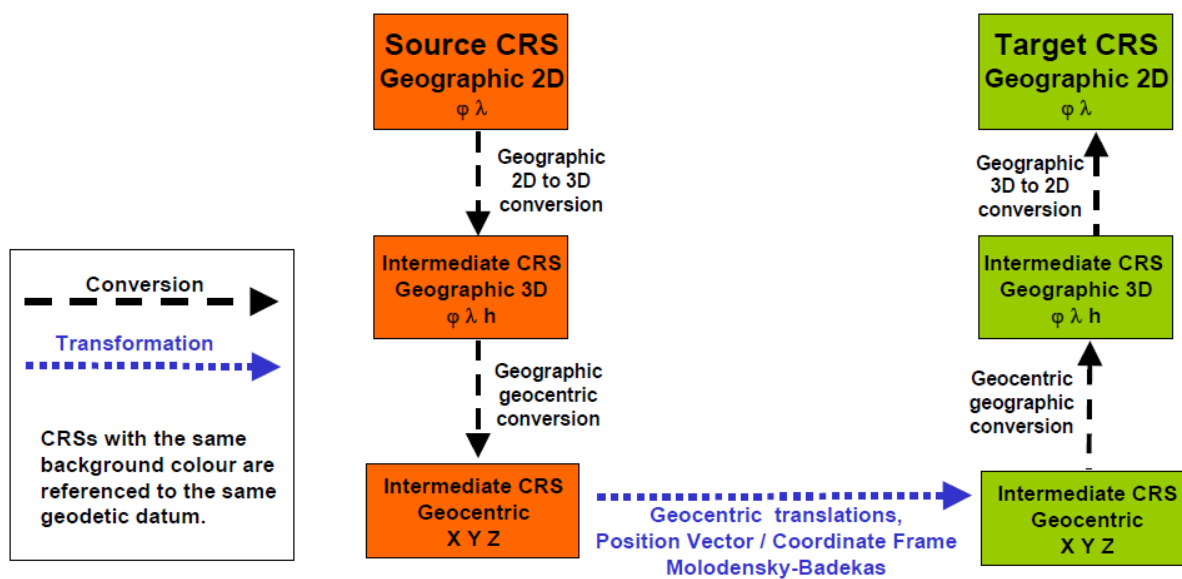


Figure 31: Implicit concatenated operation techniques with several kinds of datum transformations based on geocentric coordinates (IOGP, 2012)

## 4.2 Methods of coordinate transformations between CRSs

Often a transformation requires concatenated operations which include conversions step. The exact workflow depends on:

- the CRS domain where the transformation method operates in, and
- the type of CRS for the source and target (IOGP, 2012).

Some transformation methods operate directly between geographic coordinates, others are between geocentric coordinates. Some methods (polynomial transformation and miscellaneous linear coordinate operations) may also be encountered for use between other types of coordinate reference systems, for example directly between projected coordinate reference systems (IOGP, 2012).

The concept of implicit concatenated operations can be extended as shown in Figure 32 (CRS with the same background colour are referenced to the same geodetic datum):

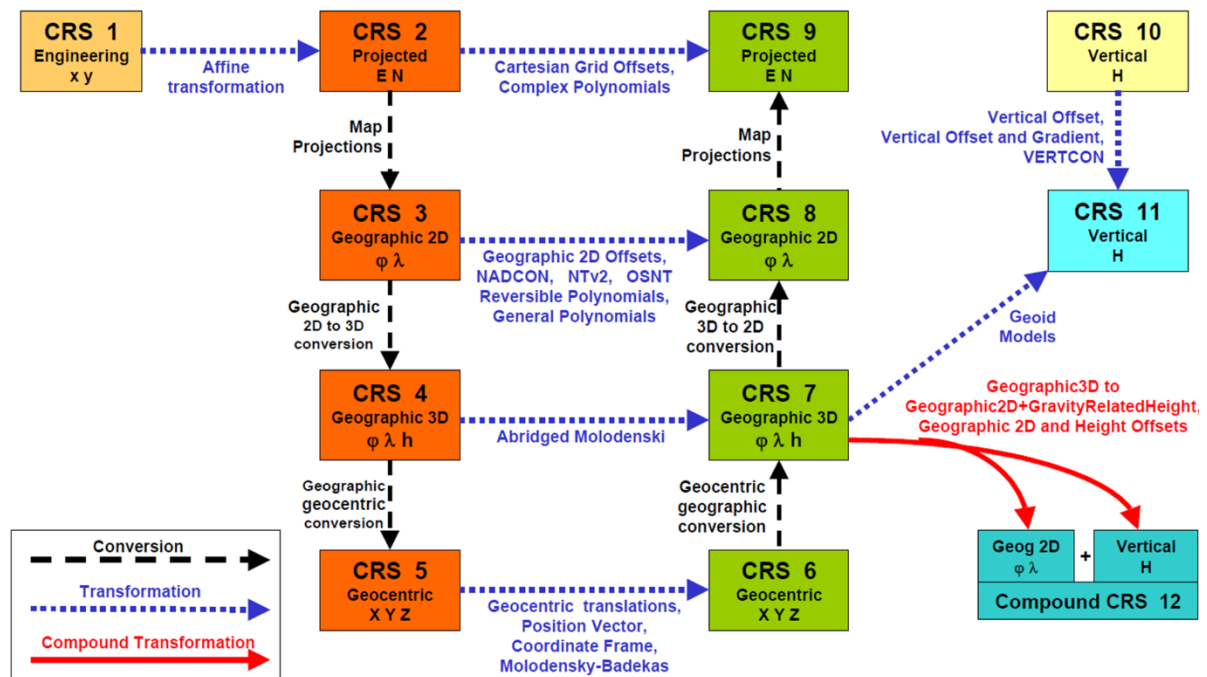


Figure 32: Implicit concatenated operation techniques (IOGP, 2012)

Transformation methods between Coordinate Reference Systems are classified into the following subtypes with respect to the kind of the source and target CRS (IOGP, 2019):

- **Geocentric Cartesian CRS:**
  - **Helmert<sup>10</sup> (7-parameter) transformation:** (three parameters for geocentric translation, three for rotation and one for scale differences), expressed in matrix form in what is known as the “Bursa-Wolf” formula. There are different definitions of rotation parameters: position vector (conform with ISO 19111:2007) and coordinate frame – they differ only in the sign of the rotation parameters (IOGP, 2019).
  - **Molodenski<sup>11</sup>-Badekas (10-parameter) transformation:** includes three additional parameters to eliminate high correlation between the translations and rotations in the derivation of parameter values for the Helmert transformation methods. Instead of being expressed about the geocentric coordinate reference system origin, they may be derived at a location within the points used in the determination (IOGP, 2019)
  - **Geocentric translation (3-parameter):** assumes that the axes of the ellipsoids are parallel, the Prime Meridian is Greenwich and that there is no scale difference between the source and target coordinate reference system (IOGP, 2019) and gives very restricted accuracy.
- **Geographic CRS:**
  - **Concatenation of three coordinate operations:** Transformation of coordinates from one geographic coordinate reference system into another is often carried out as a concatenation of the following operations: (geographic to geocentric) + (geocentric to geocentric) + (geocentric to geographic). The middle step of the concatenated transformation, from geocentric to geocentric, may be through any of the methods described above: 3-parameter geocentric translation, 7-parameter Helmert or Bursa-Wolf transformation, or 10-parameter Molodensky-Badekas transformations (IOGP, 2019).
  - **Abridged-Molodenski (5-parameter) transformation:** uses direct coordinates in geographical 3D CRSs (IOGP, 2019), additionally to the three geocentric translations there are 2 parameters to cover the different ellipsoid definitions between both CRS, and

<sup>10</sup> Helmert, Friedrich Robert (1843 – 1917): german geodesist

<sup>11</sup> Molodenski, Michail Sergejewitsch (1909 – 1991): russian geodesist

- **Geographic offsets by interpolation of gridded data:** a grid file<sup>12</sup> with given offset values for latitude and longitude (geographic 2D) in a regular grid.
- **Projected CRS:** These coordinate operation methods do not readily fit to the ISO 19111 classification of being either a coordinate conversion (no change of datum is involved) or a coordinate transformation (IOGP, 2019). They work with two projected CRSs.
  - Complex polynomial transformations,
  - Similarity or affine transformation, and
  - Cartesian grid offsets.

The 7-parameter Helmert transformation between two geodetic CRSs with geocentric 3D-coordinates is commonly used in surveying and GIS applications. In recent years, the method NTV2 for transformation directly dealing with geographic 2D coordinates of two CRSs is used more frequently. In small-scale cartography, the geocentric translation and the Abridged Molodenski transformation dealing with geographic coordinates is often used.

### 4.3 Practical notes

The accuracy level of a transformation between two CRS with different geodetic datums depends on the transformation parameters. Commonly there are various sets of parameters for one transformation which differ in accuracy and therefore in the kind of application (e.g. surveying, large scale mapping, small scale mapping).

Very often a set of parameters is valid in a specific region only (e.g. country, small area for specific surveying). A reason for this is that most national CRS are not homogeneous in themselves. Through historical processes (e.g. outdated measuring procedure and equipment or repeated adjustments) the relative accuracy of the national geodetic network is often less than it can be realised with modern measuring methods. Transformations can be calculated using high-level GIS (e.g. ArcGIS) or special geodetic programs.

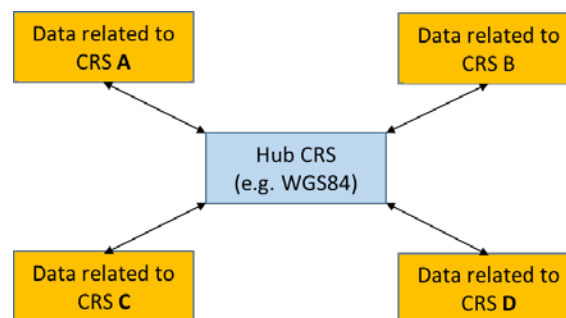


Figure 33: Hub concept in which a standard CRS is selected as a hub (based on IOGP, 2012).

International registers, for example the EPSG database, list not only descriptions of coordinate systems but also parameters for transformations between a local and the global system according to the hub system (Figure 33). Using the ArcGIS software, a standard CRS can be preset for the data frame. Thereafter, data from other CRSs may be processed if a valid transformation is provided by ArcGIS or known to the user. However EPSG transformation often result in limited accuracy. National surveying and mapping agencies distribute valid parameters for transformations for their territory which provide higher accuracy level.

## 5 Surveying with GNSS

### 5.1 GNSS introduction

**Global Navigation Satellite System (GNSS)** is the standard generic term for satellite navigation systems that provide autonomous geo-spatial positioning with global coverage. A GNSS allows small electronic receivers to more or less accurately determine their location (longitude, latitude and altitude) using time signals transmitted along a line of sight by radio from satellites. Receivers on the ground with a fixed position can also be used to calculate the precise time as a reference for scientific experiments.

<sup>12</sup> Grid file computed in a unique procedure (e.g. NTV2) from identical points, to overcome the lack of inhomogeneous transformation parameters within the validity area of the CRS.

NAVSTAR GPS was the first fully operational GNSS and Global'naya Navigatsionnaya Sputnikova Sistema (Russian Global Navigation Satellite System, GLONASS) is the second system. Some more GNSSs will be fully completed in the next few years: the European Galileo system and the Chinese BeiDou Navigation Satellite System (BDS), as well as two regional systems: Indian Regional Navigation Satellite System (IRNSS/NavIC) and Japanese Quasi-Zenith Satellite System (QZSS).

It may surprise us that so many systems are being installed for the seemingly same tasks. However, apart from the military reasons, there are increasingly economic motives, as Guenter W. Hein, former Head of ESA's EGNOS and Galileo Evolution Department explains: "One may wonder whether we need so many systems, yet, who wants to stay on the sidelines in this high-tech area? Satellite navigation with its precise positioning, navigation, and timing (PNT) information is an enabling technology and an important factor in the economic impact of new applications. Precise satellite navigation time is used in the critical infrastructure (telecommunications, power supply, etc.) of many countries, and restricted and encrypted PNT services are a major element in governmental tasks and military applications." (Guenter W. Hein in: Teunissen, Montenbruck, 2017).

The typical accuracy of single-constellation *single point positioning* (SPP) based on civilian satellites signals (for GPS called C/A-Code) is in the order of 10 m. This accuracy is basically due to the uncertainty in the orbits, satellite clocks, atmospheric delays and multipath signal transmission effects. The error sources may vary with time or location. Several approaches were developed to enhance GNSS positioning accuracy and integrity.

The *differential or relative GNSS* (DGNSS) method has already been applied for decades to improve the positioning accuracy through eliminating or significantly removing errors that are common for receivers simultaneously tracking data of the same GNSS satellites. All errors determined at a so-called reference station are merged into one parameter describing the total influence on the observation. The main problem arises from the distance-dependent errors. As the distance from the users rover receiver to the reference station increases, the validity of the correction decreases. Non-integer fixed ambiguities in the carrier phase measurement are the result.

In recent years, networks of continuously operating GNSS reference stations (CORS network) are set up to provide carrier phase corrections for precise real time positioning, also for kinematic GNSS applications (*Real Time Kinematic or RTK-GNSS*). Nevertheless, more and more absolute *precise point positioning* (PPP) models are moving into focus. The idea is to use the network of reference stations for parameterization of the individual error influences according to the state-space approach. If the state of each GNSS error component can be modelled the estimation of ambiguities will be faster and more reliable even over longer interstation distances. These PPP approaches promise more accurate position determination in regions without RTK infrastructure.

## 5.2 Introduction to NAVSTAR GPS

The NAVigation System with Time And Ranging (NAVSTAR) Global Positioning System (GPS) is an all-weather, space-based navigation system, which was designed primarily for the United States Department of Defense (DoD). Developed since 1973, it became fully operational in 1994, allowing the worldwide and instantaneous determination of a vehicle's position and velocity (i.e. navigation) as well as the precise coordination of time. Since May 2000, the earlier selective availability, an artificial degradation of the accuracy is disabled. As a result, accuracies in the range of about 10 m are possible with simple receivers. In several steps, GPS is modernized, especially by the following development:

- second civilian code on a second carrier for reducing the influence of the ionosphere signal delay,
- third carrier in L-Band and
- stronger signals.

The NAVSTAR-GPS, like any other GNSS, consists of three major segments (Figure 33):

- The **Control Segment** (also referred as ground segment) with ground based equipment for monitoring the satellites and updating the information they transmit. As its name suggests, the Operational Control System (OCS), maintains and supports the rest of the system. It has three main activities – tracking, prediction, and uploading – and consists of a single Master Control Station (MCS), an Alternative Master Control Station (AMCS) as backup, 17 monitor stations (MS), and 4 ground antennas (GA) (see figure ).
- The **Space Segment** provides global coverage with four to eight simultaneously observable satellites above 15° elevation. This is accomplished by having satellites in six nearly circular orbits with an altitude of about 20200 km above the Earth and a period of approximately 12 hours. The number of operational satellites is 21, (plus three additional active spares), with an inclination of 55° and with four satellites per plane. The spare satellites are used to replace any malfunctioning "active" satellites.
- The **User Segment** comprising an unlimited number of receivers which receive the satellite signals and calculate instantaneous position and other navigation information.

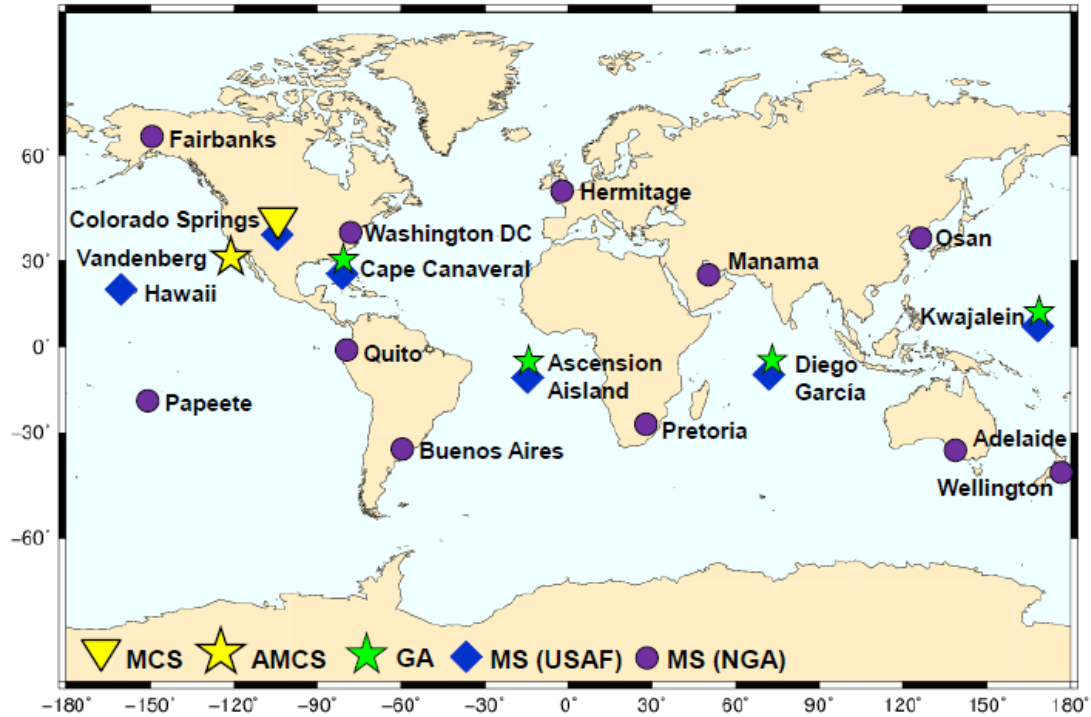


Figure 34: Infrastructure of the ground segment of GPS (Sanz Subirana et al., 2013)

## 5.3 How GPS works

### 5.3.1 Ranging Measurements

#### Signal structure

The actual carrier broadcast by the satellite is a spread-spectrum signal that makes it less subject to intentional (or unintentional) jamming. The spread-spectrum technique is commonly used today by such diverse equipment as hydrographic positioning ranging systems and wireless Local Area Network (LAN) systems. The key to the system's accuracy is the fact that all signal components are precisely controlled by atomic clocks. Rubidium or cesium atomic frequency standards or even hydrogen masers are used in GNSS satellites for this purpose. These highly accurate frequency standards at the heart of GPS satellites produce the fundamental *L*-band frequency of 10.23 MHz. Coherently derived from this fundamental frequency are two signals, the *L*<sub>1</sub> and the *L*<sub>2</sub> carrier waves generated by multiplying the fundamental frequency by 154 and 120, respectively, thus yielding

$$L_1 = 1575.42 \text{ MHz (19 cm)} \quad L_2 = 1227.60 \text{ MHz (24 cm)}$$

These dual frequencies are essential for the elimination of the major source of error, ionospheric refraction. The pseudo-ranges that are derived from measured travel time of the signal from each satellite to the receiver use two pseudo-random noise (PRN) codes that are modulated (superimposed) onto the two base carrier waves (Figure 35).

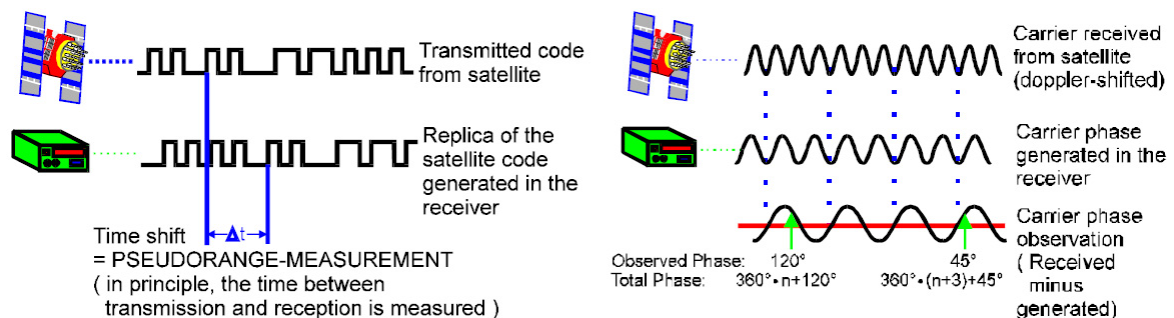


Figure 35: Measurement principles of GPS

The first code is the **C/A-code** (Coarse/Acquisition-code), also designated as the Standard Positioning Service (SPS), which is available for civilian use. The C/A-code, with an effective wavelength of 293.1m, is modulated only on  $L1$  and is purposely omitted from  $L2$ .

The second code is the **P-code** (Precision-code), also designated as the Precise Positioning Service (PPS), which has been reserved for use by the U.S. military and other authorised users. The P-code, with an effective wavelength of 29.31m, is modulated on  $L1$  and is purposely omitted from  $L2$ .

### Pseudorange

The distance from a single satellite is established by measuring the travel time of a radio signal (harmonic electromagnetic wave) from the satellite to the receiver. This can be obtained by tracking the pseudorange noise code (PRN code) modulation of the signal as illustrated in Figure 36.

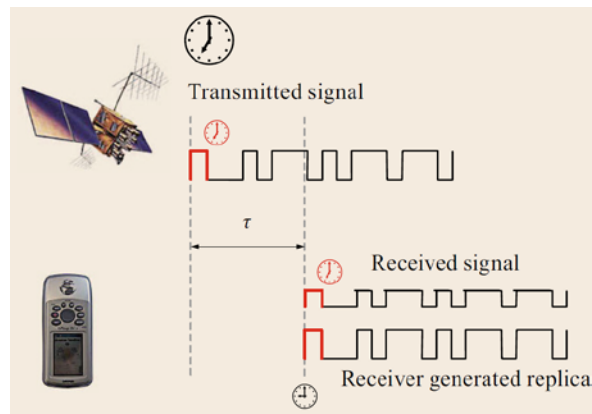


Figure 36: Princip of pseudorange measurements (Langley et al., in: Teunissen and Montenbruck, 2017)

The receiver identifies each GPS satellite based on its specific code structure. Within the receiver, a local copy of the PRN code (replica code) is generated, which is continuously compared and aligned with the signal received from the satellite. The code shift amount (also called code delay) corresponds to the time difference between the transmission and reception time. This tracking loop provides continuous measurements of the instantaneous code phase and hence the transmission time corresponding to the currently received signal. By comparing this time with the local receiver time, the signal propagation time are obtained. This time difference multiplied by the speed of light (299,792,458m/s) gives the distance, called pseudo-range.

The use of a code is important because it allows the receiver to make the comparison at any time. It also means that many satellites can operate at the same frequency because each satellite is identified by its own pseudo-random number (PRN) code. This applies to all GNSS except GLONASS, which uses different carrier signals to identify the satellites.

The modernised GPS with the new  $L5$  signal, is expected to provide civil users with horizontal accuracies of about 2-3m for 95% of the time, as good as PPS users currently enjoy, because the possibilities to reduce the influence of the ionosphere through evaluation of civil codes signals on two carriers (Sanz Subirana et al., 2013). Chapter 5.8.1 presents the renewed GPS signals.

### Basic types of measurements

Overall, the GNSS signals enable three basic types of measurements (Langley et al., in: Teunissen and Montenbruck 2017):

- Pseudorange: As stated above, this is the timing based on PRN codes. Due to the clock synchrony and other signal influences, the precision is only in the dm-range.
- Carrier Phase: A measure of the instantaneous beat phase and the accumulated number of zero-crossings obtained after mixing with a reference signal of the nominal frequency. Changes in carrier phase over time reflect the change in pseudorange but are substantially (about 2 orders) more precise. This measurement is a subdivision of a wavelength of the signal and the integer number of additional cycles, making up the remainder of the distance, which is unknown. The integer cycle count is not observed but counted by the receiver. Every loss of lock (signal interruption) leads to a loss of the number of cycles and produces a so-called cycle slip. Thus, since the initial value of  $n$  (and the one after a cycle slip) is unknown, phase measurements are ambiguous: This ambiguity (= integer number of cycles) has to be determined in the processing. Solving the ambiguities is more time and hardware consuming because of more sophisticated algorithm that require uninterrupted signal for a prolonged period of time.

- Doppler: The change in the received frequency caused by the Doppler effect is a measure of the range-rate or line-of-sight velocity.

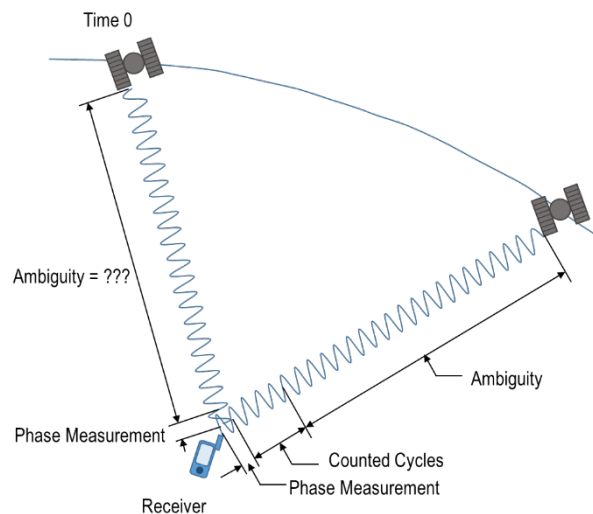


Figure 37: Principle of carrier phase measurement and its ambiguity problem

### 5.3.2 Satellite positioning

In addition to the PRN codes a data message is modulated onto the carriers consisting of:

- satellite ephemerides,
- ionospheric modelling coefficients,
- status information,
- system time and satellite clock bias, and
- drift information.

The control segment is responsible for the proper operation of the GNSS. Its basic functions are to control and maintain the status and configuration of the satellite constellation, to predict ephemeris and satellite clock evolution, to keep the corresponding GNSS time scale (through atomic clocks), and to update the navigation messages for all the satellites.

Orbit (ephemerides), clock offsets and clock drifts with respect to system time of the individual GNSS satellites are important parts of the broadcast navigation message. They are essential for the position determination.

The satellites of a GNSS are to be regarded as reference points. The receiver calculates the satellite coordinates from the broadcast message (ephemerides). GNSS broadcast ephemerides are linked to the position of the satellite antenna phase centre in the specific GNSS reference frame. Each GNSS uses an own reference system:

- GPS: World Geodetic System 1984 (WGS-84), EPSG: 4326.
- GLONASS: Parametry Zemli 1990 (PZ-90), EPSG: 4923.
- Galileo: Galileo Reference Frame (GTRF).
- BeiDou: China Geodetic Coordinate System 2000 (CGCS2000), EPSG: 4490.

The position generated by a multi-GNSS receiver using multiple systems refers to WGS-84. The transformation between the reference systems has to be solved by the receiver with known transformation parameters.

### 5.3.3 Satellite trilateration

Coordinates are calculated for any position on Earth by measuring the distances from a number of satellites to the position – the satellites act as precise reference points. If the distance from one satellite is known, the position can be narrowed down to the surface of a sphere surrounding that satellite (Figure 38).



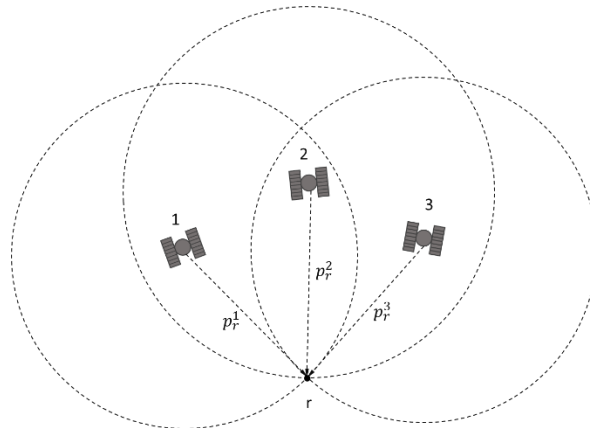


Figure 38: Principle of satellite trilateration for positioning (explained in 2D)

If the distance from a second satellite is also known, this narrows the position down to the intersection of the two spheres. Add a third satellite and the position is narrowed down to one of two points. One of these positions is disregarded – it will be far out in space or moving at high speed, and so by eliminating this position the correct position can be found. Although three satellites can be used to calculate the coordinates for a position, a fourth satellite is needed to solve the four unknowns,  $x$ ,  $y$ ,  $z$ , and time (Figure 38).

#### 5.3.4 Accurate timing

The calculations of pseudo-range using GPS depend on highly accurate clocks. The generation of all satellite codes is based on the fundamental frequency of the existing clocks. However, very different clock systems are available for this, high-precision atomic clocks in the GNSS satellites and simple quartz clocks in the GPS receivers for generating the replica codes.

Satellites have atomic clocks that are accurate to a nanosecond (one billionth of a second), but these are too expensive to put in every ground receiver. However, despite this high stability satellite clocks accumulate some offsets along time. The satellite clock offsets are continuously estimated by the Ground Segment and transmitted to the users to correct the measurements.

#### 5.3.5 Correcting errors with broadcasting parameters for global correction models

Some sources of error in GPS are difficult to eliminate. The calculations assume that the GPS signal travels at a constant speed (the speed of light). However, the speed of light is constant only in a vacuum. Once the GPS signal enters the ionosphere (a band of charged particles 80 to 120 miles above the surface of the Earth) and the troposphere, the signal slows down, resulting in incorrect distance calculations. Due to the parameters for global correction models transmitted in the navigation message, it is only possible to partially reduce these influences, e.g. for the ionospheric activity only about half.

The atomic clocks errors are very minor and are adjusted by correction terms which are calculated by the system operator and transmitted by the satellites. Satellite orbit errors can occur. The evaluations of scientific collaborative projects show that the orbital data could be predicted more accurately than the system operator does. However, satellite orbit and clock errors can be eliminated by differential methods.

#### 5.3.6 GPS for absolute positioning

Figure 39 shows that if a 3D-position is to be determined four pseudo-range measurements to different satellites have to be measured. The extra measurement is to determine the clock offset between the very precise caesium clock of the satellite and the non-precise quartz clock of the receiver. A system of equations with four unknowns (Figure 39) has therefore to be solved:

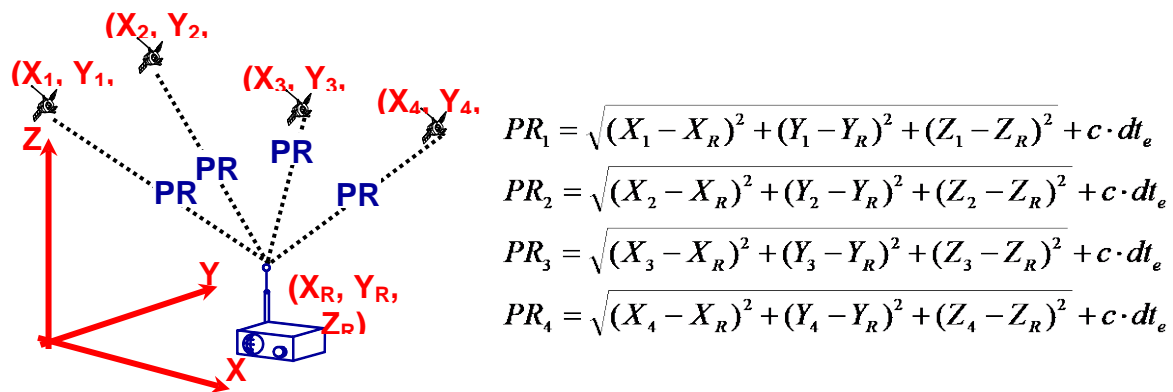


Figure 39: Pseudo-range measurements (Resnik, B., Bill, R. (2018))

## 5.4 Restriction of accuracy and techniques to avoid this

The accuracy of GNSS depends on the type of measurement (code, carrier-smoothed code, carrier), the measurement method and various other sources of error: orbits and clocks, signal propagation in ionosphere and troposphere, multi path and receiver noise. The elimination of some errors in single point positioning can be achieved by forming ‘differences’ between observations.

The positioning of a static or roving user relative to a fixed reference station with known WGS84 coordinates, is called *Differential GNSS Positioning (DGNSS)*. Here, differences between the observations of two stations are formed. Such a reference station can be used for an infinitely large number of users around it. When differential code corrections are used, one speaks of DGNSS, in the case of differential phase corrections they are called Real Time Kinematic GNSS (RTK-GNSS).

Errors in the satellite segment are substantially corrected by differential methods since they act equally on both stations. Theoretically, the determined differential corrections are valid only to the exact position of the reference station and the time of the generation of the correction, because some error components, such as signal propagation errors in the atmosphere are not the same size, with increasing distance between rover and reference.

However, the validity of the code corrections decreases only slowly, so that distances of a few hundred kilometers may lie to the reference station. In contrast, for carrier phase corrections, the range to the reference station may only be a few of 10 kilometers, otherwise ambiguity solutions will be more critical, especially in times of active ionosphere.

In order to improve the range in many countries there are providers who operate networks of permanent reference stations. The observations of all reference stations run to a central office, which transmits corresponding correction data from surrounding stations to the user depending on the transmitted estimated user position. This also overcomes dependence on a station and can compensate for failures. The correction calculation is optimized by spatially weighting the corrections from three neighboring reference stations. The result is a stream of correction data for a virtual reference station (VRS) located a few meters from the user, transmitted in real time via mobile internet (network RTK).

Errors occurring in the user receiver and the immediate receiver environment, like receiver noise and multipath are uncorrected by Differential GNSS since they are not equally effective on the reference station.

Multipath interference can introduce errors into a GPS position. This occurs when the signal is reflected off other objects at or near the Earth’s surface. The reflected signal interferes with the straight line signal. The deflected signal is superimposed with the direct signal and leads to an extension of the distance measurement in relation to the true distance. Its important when the signal comes from a satellite with low elevation. This error is different for different frequencies. It affects the phase measurements, as well as the code measurements. Advanced signal processing and well-designed antennas help minimize the effect of multipath.

Another error on the user side is the receiver noise, which Sanz Subirana et al. explains as “... a white-noise-like error and can be smoothed using a low-pass filter. This error affects both the code and carrier measurements, but at different magnitudes. The accuracy of pseudorange measurements is about 1% of the wavelength (‘chip’), or better. This means, for instance, noise with a maximum value of 3 m for the GPS civil C1 code (i.e. C/A code) and about 30 cm for the protected P codes. However, when smoothing the code with the carrier phase, the C1 code noise can be reduced to about 50 cm. The carrier phase noise is at the level of few millimetres (about 1% of the carrier phase wavelength).“ (Sanz Subirana et al., 2013).

## 5.5 Measurement methods

The receivers are differentiated according to the type of pseudorange distance measurement as mentioned in section 5.3.1 “Basic types of measurements” in,

- PRN code measurement which used by all standard receivers (consumer class),
- carrier phase measurement on one or more frequencies with fully ambiguity solution (geodetic receiver).

However, GNSS receivers are also be classified according to their positioning accuracy ranges, which they achieve depending on the measurement type (PRN code or phase derived pseudorange) and measurement method (absolute or differential).

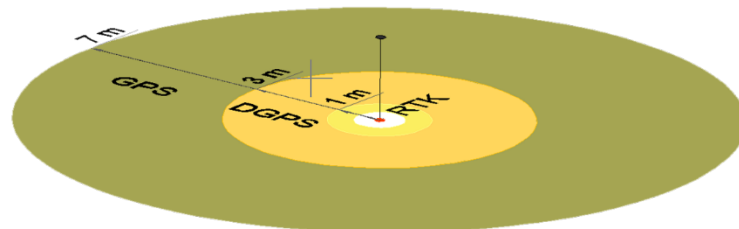


Figure 40: Comparison of accuracies achievable with different GNSS methods from several meters up to 1-2 cm.

### 5.5.1 Meter- and sub-meter accuracies

*Pseudo-range or Phase-Smoothed.* Both methods are be used with absolute GNSS or differential GNSS (DGNSS). In the case of DGNSS it depends on whether there is a correcting signal or not. Here the receiver costs are moderate (receivers with phase smoothing are ten times more expensive than standard GPS receivers but still cheaper than receivers using carrier phase measurement with complete integer solution of phase ambiguity. Of course, phase-smoothing techniques are no more accurate than phase measurement with ambiguity solution, but more robust, especially when short measurement times are required.

### 5.5.2 Centimeter accuracies

*Carrier Phase Based Approach.* The necessary equipment and analysis software has a significantly higher price, mainly also due to the use of dual frequency receivers. Different observation strategies are possible: static, rapid or fast static, pseudo-kinematic, stop-and-go, semi-kinematic and kinematic modes. Furthermore, differential mode is needed to provide dm- or cm-accuracy.

The main differences to the phase-smoothed approach are solving the ambiguity problem by the rover-receiver and the presence of true carrier phase corrections of a reference station (base station). In the case of a multi-frequency receiver, this is called Real Time Kinematic GNSS (RTK-GNSS), sometimes also referred as Precise Differential GNSS (PD-GNSS).

The RTK procedure requires the undisturbed reception of signals from at least five GPS satellites (at least six satellites are recommended). These must be the same satellites on both stations (reference and rover). Position accuracies of a few centimeters are achievable. The robustness of the ambiguities solutions depends strongly on the distance to the reference station. To reduce this dependence, reference stations are networked with each other. The observations from three reference stations are weighted for the rover's approximated position and then transmitted to the user's rover. This is done through various techniques:

- Surface correction parameter, which requires a higher volume of data and more intelligence on the rover. This technique is often referred to as FKP, an abbreviation of the German term Flächen-Korrekturparameter.
- Summarized as corrections of a single virtual reference station (VRS), which means less data volume, no further calculations on the rover. The most widely used format for transmission of real-time GNSS observation data and intermediate products is RTCM Standard 10 403-2. Positioning results are transmitted using National Marine Electronics Association (NMEA) 1083.

### 5.5.3 Post-processing

In the case of a missing data connection between reference and rover or a lack of infrastructure for reference data, a subsequent calculation of the position can be carried out. This is based on observation logs of both stations during the same period. The three-dimensional relative baseline vector, between the reference station and the user station, results from processing using GNSS analysis software.

The disadvantages are longer occupation times and the lack of certainty as to whether the observation period is sufficient to provide high accuracies. In the case of post-processing and transmission of data files, receiver independent exchange format (RINEX) is the most common open format. It is used for GNSS observations and broadcast ephemerides.

## 5.6 Accuracy management

Surveying with only one ordinary code receiver using the civil code signal without corrections of a second receiver acting as reference, does not provide an independent means of assessing the currently achievable accuracy. Only estimates of the position accuracy are possible, which can be based on the *number of satellites*, the measure of the *quality of the satellite distribution*, the *signal-to-noise ratio* of the received signals and an *evaluation of the overdetermined position calculation*. Some receiver shows an estimated value for position error based on the factors mentioned above. How this is derived, is usually not disclosed by the manufacturer.

### 5.6.1 Quality of satellite distribution

In addition to a large number of satellites, the geometry of the satellites (i.e. how the user sees them) affects the positioning error. Unfavorable constellations occur under dense trees, in narrow streets or valleys in mountains.

The size and shape of the uncertainty region caused by measurement noise change depending on the intersection angle caused by different satellite positions. This effect is called **Dilution Of Precision (DOP)**. DOP factors are the results of a calculation that takes into account each satellite's location relative to the other satellites in the constellation.

A low DOP indicates a higher, a high DOP indicates a lower probability of accuracy.

The DOP factor has a geometric equivalent: the size of the volume of the body spanned by the satellites and the receiver antenna. The optimal distribution of five satellites is achieved when four satellites are distributed flat over the horizon at 90 degrees and the fifth satellite is at zenith.

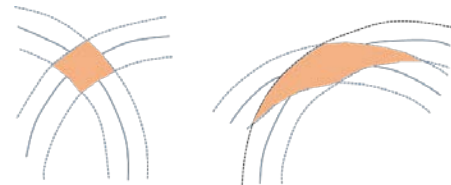


Figure 41: Dilution of precision effect

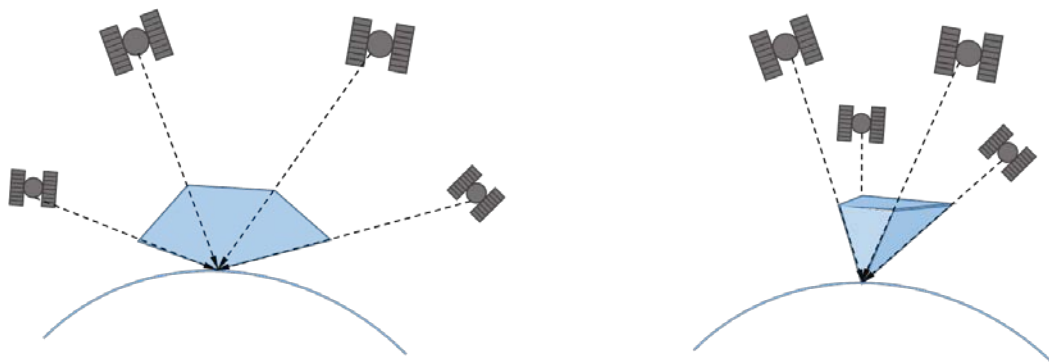


Figure 42: Relation between satellite distribution and the position uncertainty represented over the body volume. (Left figure: good satellite distribution/good intersection angles = large volumes = low DOP. Right figure: worse satellite distribution/worse intersection angles = low volume = high DOP.)

You can select the best data collection time based on reports and graphs showing times of lowest DOP. Planning software can predict the geometric, position, horizontal, vertical, and temporal (clock offset) dilutions of precision separately (PDOP, HDOP, VDOP, and TDOP respectively):

- **Time (TDOP)** refers to clock offset.
- **Vertical (VDOP)** refers to altitude.
- **Horizontal (HDOP)** refers to horizontal measurements (latitude and longitude).
- **Position (PDOP)** refers to position measurements (latitude, longitude, and altitude).
- **Geometric (GDOP)** refers to the position and the time

These predictions do not take into account local obstructions unless you enter these obstructions into the software. Obstructions might block some lower-elevation satellites; if such satellites were included in the DOP computation, the computation would be incorrect.

In some GNSS receivers, a threshold can be set so that the user receives warning about high DOP values. The HDOP can be used instead of the PDOP and is useful if you are more concerned with the horizontal accuracy of your data than with vertical accuracy. A HDOP of one is the optimum, a HDOP of three or below gives excellent positions. A HDOP of five or more is poor.

**Summary:** The DOP indicates as a factor by how much the error of the distance measurement increases, by the influence of the satellite distribution. An HDOP value of about one is the optimum. A HDOP value of three means the same distance measurement accuracy is three times worse due to the less favorable satellite distribution. In times when multi-constellation GNSS receivers are used, attention to the satellite distribution is not important if there is little or no signal blocking around the horizon. In contrast, the DOP is important when dealing with GNSS measurements in environments characterized by strong signal obstacles (urban environment, forest, mountains).

### 5.6.2 Quality of satellite signal reception

A good signal-to-noise ratio is inherent to the correct signal reception. The reception can be adversely affected by obstacles in the signal path, for example under trees. It is also often unfavorable for low elevation satellites above the horizon. The quality of the phase measurement with the geodetic receivers depends even more on an undistorted and stable carrier phase signal. Therefore, some receivers allow a presetting of a so-called *elevation mask* or *cut-off elevation angle*: satellites with an elevation angle smaller than the predefined cut-off angle will not be taken into account.

### 5.6.3 Quality of differential correction signal reception

Differential GNSS methods are based on temporal and spatial valid corrections of one or more reference stations. In addition to the spatial validity, and thus refers to the similarity of the errors at both stations, the corrections received from the rover may have only a low latency.

Especially with RTK, low latencies for corrections of a few seconds are necessary (near real time).

As shown in Figure 43, when measuring with geodetic receivers, different quality parameters can be displayed to be aware of their values in the field:

- Theoretical number of satellites currently available at the site.
- Number of received satellites separated for the different carriers *L1* and *L2* (now also *L5*).
- Accuracy of the position here displayed as a symbol (here high accuracy, shown by position cross with a small error circle).
- Latency of RTK correction data (RTCM data).
- DOP factors of the distribution of the satellites used for position calculation.

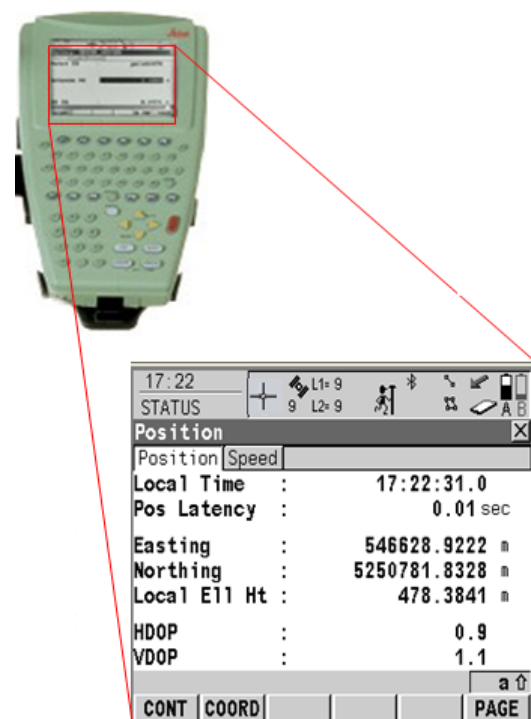


Figure 43: GNSS receiver with displayed position and parameter for assessing the accuracy.

## 5.7 GNSS augmentation systems

Augmentation of a Global Navigation Satellite System (GNSS) is a method of improving the navigation system's attributes, such as accuracy, reliability and availability, through the integration of external information into the calculation process. There are many such systems in place and they are generally named or described based on how the GNSS sensor receives the external information. Some systems transmit additional information about sources of error (such as clock drift, ephemeris or ionospheric delay), others provide direct measurements of how much the signal was off in the past, while a third group provide additional vehicle information to be integrated in the calculation process.

### 5.7.1 Satellite-based augmentation system (SBAS)

A Satellite-based Augmentation System (SBAS) is a system that supports global or wide-area (regional) augmentation through the use of additional satellite-broadcast messages transmitted in real time to the user's GNSS equipment. Such systems are commonly composed of multiple ground stations building up a network, located at accurately-surveyed points. The ground stations take measurements of all visible navigation satellites of one or

more GNSSs/RNSSs. The resulting calculations at the master control station are then broadcast over the covered area using geostationary satellites that serve as an augmentation, or overlay, to the original GNSS message.

The SBAS network provide differential correction data and integrity information about the augmented GNSSs/RNSSs. The systems use a state-space-domain approach in which corrections for specific error sources are determined, e.g. GNSS satellite orbit and clock data, ionospheric propagation delays. Data distribution is performed via L1-bandwidth signals (1575.42 MHz) according to international standards.

User GNSS receiver performs overlapped processing of this correction data and originally GNSS signals, which allows solving navigational tasks with improved precision and reliability characteristics. In comparison, "normal" GPS receivers generally offer about 10 m accuracy, and ones using SBAS improve this to about 3 m.

Depending on the extent of the operation area of the correction services, a distinction is made between global, such as Omnistar and StarFire (both commercial providers), or regional services, such as WAAS (North America) and EGNOS (Europe).

While SBAS designs and implementations may vary widely, with SBAS being a general term referring to any such satellite-based augmentation system, under the International Civil Aviation Organization (ICAO) rules a SBAS must transmit a specific message format and frequency which matches the design of the US Wide Area Augmentation System.

Therefore, the commercial proprietary systems such as Omnistar and Starfire are not SBAS according to the ICAO standard. But in contrast, Omnistar and Starfire services provides additionally to the corrections for the code signals also corrections for dual-frequency carrier signals to obtain absolute position accuracy worldwide under 10 cm with dual-frequency GPS rover receiver.

Currently, several SBASs are in full operation or about to be completed:

- the FAA's WAAS,
- the European Geostationary Navigation Overlay Service (EGNOS),
- Japan's Satellite-based Augmentation System (MSAS), and
- India's GPS-aided GEO Augmented Navigation System (GAGAN),

In some other regions of the world further SBAS should also be in development or under study: (Russia (SDCM), South Korea (KASS), Africa (AFI), South/Central America and the Caribbean SBAS (SACCSA) and Australia. The most advanced of these projects at the moment is the Russian SDCM (ESA Navipedia, 2019).

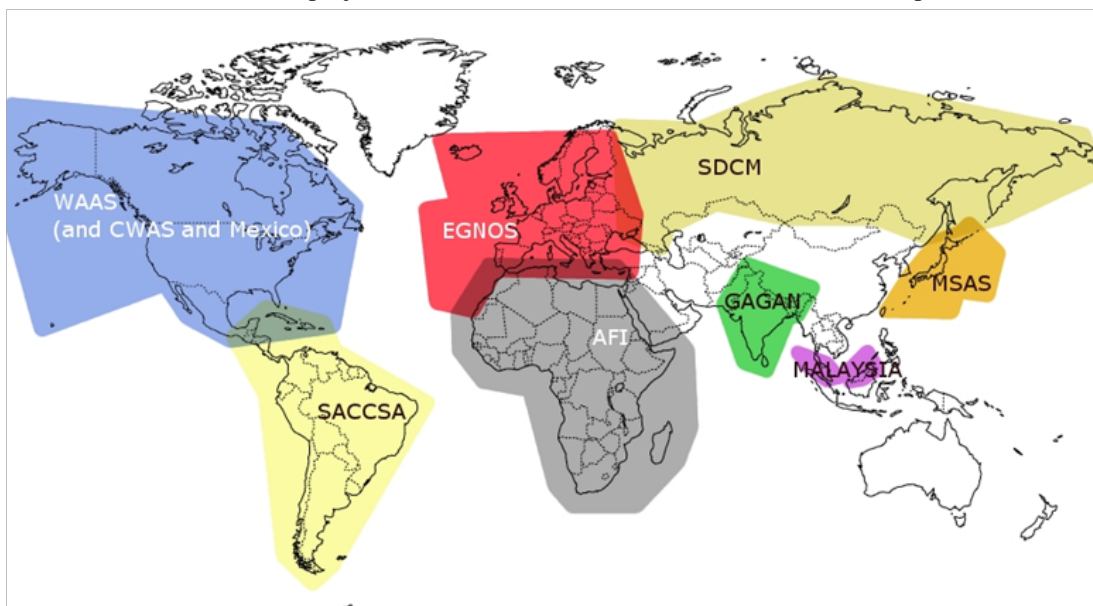


Figure 44: SBAS systems either operational, under development or under study (ESA, 2019)

Today's SBAS only augment GPS on L1 signal, with the exception of the Russian SDCM, which will augment GPS and GLONASS.

The expansion to further civil signals for GPS (multi signal), implementing the dual or multi frequency (multi frequency) and extension of the service to augment GLONASS or other GNSS is to be expected (multi GNSS). According to the Interoperability Working Group (IWG) of SBAS System the aforementioned evolutions are expected in the coming years, so a seamless navigation will be possible from and to any two locations in the world.

### 5.7.2 Ground-based augmentation system and ground-based regional augmentation system

Ground-based augmentation system (GBAS) and ground-based regional augmentation system (GRAS) describe a system that supports augmentation through the use of terrestrial radio messages. As with the SBAS described above, ground based augmentation systems are commonly composed of one or more accurately surveyed ground stations, which take measurements concerning the GNSS, and one or more radio transmitters, which transmit the information directly to the end user.

Generally, GBAS networks are considered to be localised, supporting receivers within 20km, and transmitting in the Very High Frequency (VHF) or Ultra High Frequency (UHF) bands, whereas GRAS is applied to systems that support a larger, regional area and transmit in the Low frequency (LF) and Medium frequency (MF) bands. Through the successful development of infrastructure frameworks such as NTRIP for network-transported GNSS data, GNSS corrections are increasingly being transmitted over NTRIP-based mobile Internet connections.

## 5.8 Comparison of the GNSS

### 5.8.1 The renewed GPS

For a long time the NAVSTAR Global Positioning System (GPS) was the only fully functional, fully available global navigation satellite system. It consists of about 30 medium Earth orbit satellites in six different orbital planes, with the exact number of satellites varying as older satellites are retired and replaced. Operational since 1978 and globally available since 1994, GPS is currently the world's most utilised satellite navigation system.

GPS is renewed for many years. Already with the satellites of the series Block IIR-M in the years 2005 to 2009 new civil and military signals were put into operation. The implementation of a number of advanced signal design features is continued with block IIF and current block III. A third carrier frequency was introduced and is called *L5* (1227.60 Mhz).

Today, a distinction is made between legacy and modernized signals. For all of the modernized civilian signals, these features include dataless components, at least 10 times longer PRN codes, and various improvements to the navigation data encoding and content. *L5* and *L2C* additionally employ secondary codes, and *L5* and *L1C* use wider bandwidth modulations (Christopher J. Hegarty in: Teunissen, Montenbruck, 2017).

- Legacy signals:
  - C/A-code on *L1*-Band
  - P-code on *L1*-Band (encrypted P(Y)-code)
  - P-code on *L2*-Band (encrypted P(Y)-code)
- Modernized signals:
  - L1C-Code on *L1*-Band
  - M-Code on *L1*-Band
  - L2 CM on *L2*-Band
  - L2 CL on *L2*-Band
  - M on *L2*-Band
  - I5 on *L5*-Band
  - Q5 on *L5*-Band

In case of *L5*, both the carrier and the signal are referred to as *L5*. The new *L5* code signal will be interoperable with that of Galileo, QZSS, and IRNSS/NavIC. *L5* have an improved power level and a wider bandwidth. The *L1C* is backward compatible with the current civil signal on *L1* (*L1*-C/A), and work with a higher power level and an advanced code design for enhanced performance.

It is to be expected that in the next generation of GPS receivers the processing of four civilian codes usable on three carriers will be implemented: C/A-L1, L1C, L2C and L5. Soon techniques will be used which also allow for code signals the reduction of the ionosphere signal delay by multi-frequency processing methods. As a result, significantly increased accuracies for positioning are possible.

### 5.8.2 GLONASS

Next to GPS, GLONASS is the second fully operational GNSS. The formerly Soviet, and now Russian Federation, global navigation system was initiated in the 1980s and is called GLONASS. The system first achieved its full operational capability in 1995. Following a temporary degradation, the nominal constellation of 24 satellites was ultimately reestablished in 2011 and the system has been in continued service since then.

Similar to its US counterpart, the NAVSTAR GPS, GLONASS provides two types of services:

- An open service with unencrypted signals in up to three frequency bands (*L1*, *L2*, and recently *L3*) that is globally available for all users without any limitations.

- A service for authorized users, using encrypted signals in presently two frequency bands ( $L1$ ,  $L2$ ).

“Unlike GPS, the signals of the authorized service are presently not encrypted. Even though their structure and data contents have not been publicly released by the system providers, the employed ranging code has, nevertheless, been revealed already in the early days of GLONASS through a systematic code search. This has enabled the design of geodetic dual-frequency GLONASS receivers and allowed for an early use of GLONASS in precise point positioning applications.” (Revnivykh et al., in: Teunissen, Montenbruck, 2017). Revnivykh points out that at any time an access barrier can be introduced by the Russian Federation Defense Ministry and therefore its unofficial use should be considered with due care.

The GLONASS satellites have no resonance with rotation of the Earth, so each eight days a satellite passes over the same point on the Earth’s surface. The orbital inclination of the GLONASS satellites (about 65 degree) is roughly ten degrees higher than that of other medium altitude Earth orbit (MEO) navigation systems (GPS, BeiDou, Galileo), which provides a reduced visibility gap around the celestial pole (Revnivykh et al., in: Teunissen, Montenbruck, 2017).

In the years 2012-2020 GLONASS went through a phase of modernization aimed at continuous performance increase through improvements to the ground and space systems. “With the first GLONASS-K1 satellite launched in 2011, GLONASS started to transmit additional code division multiple access modulation (CDMA) signals on the new  $L3$  signal. As part of the ongoing GLONASS modernization, CDMA signals will also be transmitted in the  $L1$  and  $L2$  bands to improve interoperability with other GNSSs, specifically GPS.” (Revnivykh et al., in: Teunissen, Montenbruck, 2017). Until then, GLONASS was the only one of the GNSS that does not use the CDMA method to identify the satellites over a unique code sequence, instead it uses the FDMA (frequency division multiply access). In the FDMA method, all satellites uses the same ranging code, but splits the L-band into slightly different frequencies so that each satellite transmits on its own frequency and thus allow concurrent processing in the receiver.

### **5.8.3 Galileo**

Galileo is tasked by the European Union (EU) as a standalone, worldwide available and independent of other non-civilian GNSS. Galileo is nearing completion, a few years from its planned completion. It is designed to be compatible with all existing and planned GNSS and interoperable with GPS and GLONASS. Galileo is operated by the European GNSS Agency (GSA). As usual in the EU, the responsibilities are spread over several countries and offices. So there are two control centers, which have divided the operator tasks, but can also work independently in the case of backup.

Two key aspects could distinguish Galileo over the other GNSS: Firstly, it is the only civil satellite navigation system under democratic control and, secondly, the positional accuracy that can be achieved is higher.

The launches of the last four of a total of thirty Galileo satellites are scheduled for the end of 2020. The system is already in operation, and this year the construction of the so-called high-precision service is to begin, which should enable more exact position determination around the globe than rival systems. The inclination of the orbit planes provides for better coverage in the higher latitudes, for example when compared to GPS.

The Galileo system, once fully operational, will offer four high-performance services worldwide (GSA, 2018):

- Open Service (OS): Galileo open and free of charge service set up for positioning and timing services. The OS comprising the data-pilot pairs E1-B/C, E5a-I/Q and E5b-I/Q, representing the publicly accessible positioning service.
- High Accuracy Service (HAS): A service complementing the OS by providing an additional navigation signal and added-value services in a different frequency band. The HAS signal can be encrypted in order to control the access to the Galileo HAS services.
- Public Regulated Service (PRS): Service restricted to government-authorized users, for sensitive applications that require a high level of service continuity.
- Search and Rescue Service (SAR): Europe’s contribution to COSPAS-SARSAT, an international satellite-based search and rescue distress alert detection system.

“Each Galileo satellite provides coherent navigation signals on three different frequencies. Each signal contains several components, comprising always at least one pair of pilot and data components.” (Falcone et al., in: Teunissen, Montenbruck, 2017).

### **5.8.4 BeiDou 2**

The People’s Republic of China is developing a separate GNSS. Three steps for constructing the system were planned (Yang et al., in: Teunissen, Montenbruck, 2017):

- The first step, was the BeiDou Navigation Satellite Demonstration System, which is called BeiDou-1 or simply BDS-1. In 2000, two BeiDou experiment satellites were launched, followed by a third in 2003.



- The second step is the regional BeiDou Navigation Satellite System. The first satellite, a medium Earth orbit (MEO) satellite was launched in 2007. Operational navigation service available for China and large parts of Asia-Pacific region was declared by the end of 2012. It is accomplished through a constellation of 14 satellites, including five satellites in geostationary Earth orbit (GEO), five satellites in inclined geosynchronous orbit (IGSO), and four MEO satellites.
- As a third step, the BeiDou Navigation Satellite System with global coverage (BDS-3) is built-up, which will be completed around 2020.

The BDS-3 will be a space constellation of 35 satellites, which include 5 geostationary orbit (GEO) satellites, which are located above the Asian region, 27 medium Earth orbit (MEO) satellites and 3 IGSO satellites which will offer complete coverage of the globe. The MEO and IGSO satellites orbits are characterized by an inclination of 55° at altitudes of 21 500 and 36 000 km. The IGSO satellites orbit the Earth in three different planes but exhibit a common ground track with its ascending node at 118°E (Yang et al., in: Teunissen, Montenbruck, 2017).

“BDS-3 will provide an open service and an authorized service in four frequency bands, including *B1* (1559–1610MHz), *B2* (1164–1219MHz), and *B3* (1240–1300MHz) with center frequencies of 1575.42MHz, 1191.795MHz, and 1268.52MHz. A new S-band signal, *Bs* (2483.5–2500MHz), is also broadcasted by the newly launched satellites. Compared to the regional BeiDou system, new and advanced signal structures are employed for better performance, compatibility, and interoperability with other GNSSs” (Yang et al., in: Teunissen, Montenbruck, 2017).

The code-only position of the free service will have a 6 meter horizontal location-tracking accuracy and 10 meter in vertical direction (Yang et al., in: Teunissen, Montenbruck, 2017).

### 5.8.5 Summary of GNSS Signals

To summarize the previous sections, Figure 45 compares the different GNSS signals and their relative allocations on the associated frequency bands.

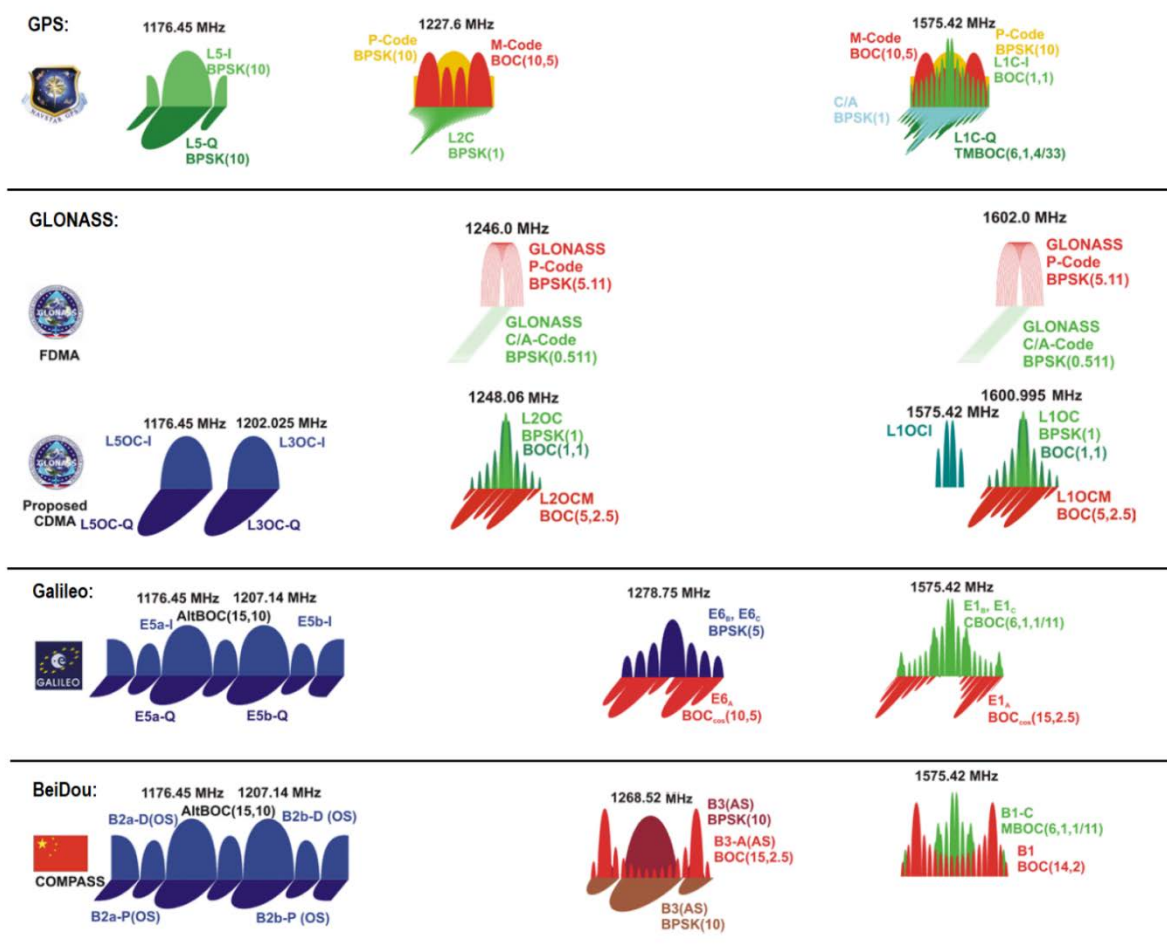


Figure 45: Overview of all GNSS frequencies and signals (source: Stefan Wallner, ESA Navipedia)

System	GPS	GLONASS	Galileo	BeiDou-3
Operator	Department of Defense, USA	Russian Space Agency and Department of Defense, Russia	Galileo Navigation Satellite System Agency (GSA), European Union (EU)	Military, China
Full Operational Capability	1995	1995 (in the 2000s, the constellation was degraded.) 2011	2020 (planned)	BDS-1: 2003 BDS-2: 2012 BDS-3: 2020 (planned)
Orbits	medium altitude Earth orbit (MEO) satellites	medium altitude Earth orbit (MEO) satellites	medium altitude Earth orbit (MEO) satellites	MEO (27 sats.) GEO (5 sats.) IGSO (3 sats.)
Number of satellites	24 normally, 31 operational	24 normally, 24 operational	22 operational, 30 planned: 24 operational +6 reserve)	33 operational, 35 planned
Constellation	6 planes, 56° inclination	3 planes, 64.8° inclination	3 planes, 56° inclination	3 planes, 55° inclination
Orbit altitude	20 350 km	19 140 km	23 222 km	21500 km (MEO) 36000 km (IGSO)
Carrier frequencies, MHz	L1: 1575.42 L2: 1227.6 L5: 1176.45	G1: 1598.0625-1607.0625 G2: 1242.9375-1249.9375 G3: 1198.55-1205.30	E1: 1575.42 E5: 1191.795 E6: 1278.75	B1: 1575.42MHz, B2: 1191.795MHz, B3: 1268.52MHz
Services	SPS, PPS	SPS, PPS	OS, CS, PRS	OS, AS, WADS, SMS
SPS: Standard Positioning Service; PPS: Precise Positioning Service; OS: Open Service; AS: Authorized Service; WADS: Wide Area Differential Service; SMS: Short Message Service; CS: Commercial Service; PRS: Public Regulated Service;				

Figure 46: GNSS Overview (Source: Langley et al., in: Teunissen, Montenbruck, 2017, Encyclopaedia Astronautica, 2019)

## 6 Position accuracy measures

### 6.1 Definitions

Ideally, a measurement device is both accurate and precise, with measurements all close to and tightly clustered around the known value. Although the terms precision and accuracy are often used interchangeably, there is an important difference between them.

Accuracy describes the closeness of value to the true or accepted value. When all values are grouped tightly together, the cluster is considered precise. They are not necessarily near the true value.

**Accuracy:** “Closeness of agreement between a test result and the accepted reference value” (ISO 3534-1).

**Positional accuracy:** “Closeness of coordinate value to the true or accepted value in a specified reference system.” (ISO 19116)

**Precision:** “Measure of the repeatability of a set of measurements” (ISO 19116). Precision “...is the closeness with which repeated measurements made under similar conditions are grouped together...” (DMA TR 8400.1) and

“...is usually expressed as a statistical value based upon a set of repeated measurements such as the standard deviation of the sample mean” (ISO 19116).

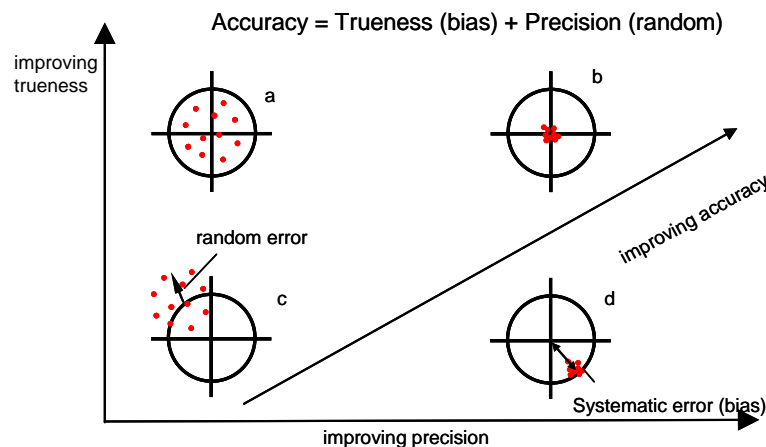


Figure 47: Accuracy and precision

Precision is affected only by the random errors in the measuring process while accuracy is affected by the precision as well as the existence of unknown or systematic errors. The difference between precision and accuracy is illustrated in Figure 47 where the plots of errors in a circular distribution are shown. In Figure 47-b, the points are grouped closely together and the measurement is said to be “precise”. It is also accurate because the centre of the group coincides with the centre of the circle. In Figure 47-d, the grouping is still precise but inaccurate because it is not centred on the centre of the circle. Instead, the mean of the points is offset by a systematic error or bias. The measurements shown in Figure 47-d are “inaccurate” because of the bias even though they are “precise” (grouped closely). In Figure 47-c, the points exhibit neither close grouping nor nearness to the centre. They are, therefore, not precise and not accurate. Measurements can be precise and inaccurate at the same time, but they can never be accurate unless they are precise at the same time Figure 47-a.

## 6.2 Horizontal position accuracy measures

### 6.2.1 Precision indexes and probability levels

“It is not always possible to remove systematic errors from positioning information. To give the user knowledge of the accuracy of the product he is using, methods have been devised to state the uncertainty of the product”. (DMA TR 8400.1). For more information see Harre (2001).

### 6.2.2 Horizontal position circular precision indexes

Horizontal position circular precision indexes deal only with the random errors in the X and Y axes. Normally, the x and y standard deviations have a different magnitude. The resulting ‘natural’ *error contour* is an ellipse with the standard deviations as semi-axes. For an easier interpretation, one would like to indicate only one value for the description of the two-dimensional error. The resulting genuine error contour is an ellipse in the general case or a circle in the special case that both standard deviations are equal. In this way, we can generate an error ellipse for any desired probability level.

Error ellipses as an error contour are not very practical for navigation, as five parameters are needed for their description: the origin, the length of each semi-axis and the orientation (e.g. angle between the major semi-axis and the x-axis). A more useful contour is a circle, which is fully described by the origin and the radius. In the case of error contours an additional parameter is the ‘confidence content’, i.e. the probability that a measured position is within the contour. For navigation a probability of 0.95 or 95% is usually required.

### 6.2.3 Root mean square (RMS)

RMS is the square root of the average of the square errors. Depending on the dimension of the error consideration, there are the vertical RMS, the (horizontal) 2D RMS and the 3D RMS.

In the case of horizontal RMS or 2D-RMS, the RMS is a single number that expresses an horizontal accuracy. This measure refers to an error radius. In order to compute the 2D-RMS of horizontal position errors, the standard errors ( $\sigma$ ) from the known position in the directions of the coordinate axis are required. 2D-RMS is the square root of the average of the square errors.

Contrary to one-dimensional statistics, there is no fixed probability level for this error measure. The probability level (p) depends on the ratio of standard deviations in the two dimensions.

Owing to the low probability content of the horizontal RMS or 2D-RMS error circle, twice the RMS of the horizontal errors is commonly required for horizontal position-finding errors:  $2D-RMS = 2drms$ , which refer to a probability level of 95 to 98 percent.

Today, technical literature still includes the error circle measures 2D-RMS and 2drms. It is important to note that 2drms is the double horizontal RMS. Both are easy to calculate, but not precise error measures since they do not contain fixed error probabilities. An indication without fixed probability should be avoided.

A more exact way is to visualise the user's position-finding accuracy, upon operator request, by a CEP95 error circle or – still better – by a 95% error ellipse. This can enhance confidence in special situations during navigation and increase the efficiency in surveying applications.

#### **6.2.4 CEP (CEP 50)**

Since the error circles DRMS and 2DRMS comprise error probabilities varying with the ratio of the coordinate error standard deviations, a procedure was developed which can be used to calculate error circles providing fixed probability content (Harter 1960; Burt et al. 1965).

„The expression CEP50 does not imply that there is a 50% probability that an error of 5m will occur, rather it means that there is 50% probability that error will not be larger than 5m.“ (DMA TR 8400.1)

The procedure is based on tabulated values that are calculated for specified probabilities and graduated ratios of the error standard deviations by evaluating the probability integral. The smaller of the two standard deviations is to be multiplied by a value taken from the table. The result of the calculation is the error circle radius with the desired probability content. Since the use of tables is not optimal for automated calculation, approximation polynomials for the tabulated values have been determined (Harre, 1987). Today, mainly the error circle CEP95 containing 95% of the position fixes of a set of measurements is of importance for navigation purposes, the calculation of which is included in the proposed algorithm (Harre, 2001).

Today, the error circle CEP95 containing 95% of the position fixes of a set of measurements is of most importance for navigation purposes (Harre, 2001).

The only measure for navigation errors that provides a fixed probability is the 'Circular Error Probability', the CEP value, to which usually an index indicating the probability is added. Other error measures, such as DRMS or 2DRMS do not provide a predefined probability (Harre, 2001).

## **7 Mobile GIS**

### **7.1 Introduction**

Mobile GIS is the expansion of a geographic information system (GIS) from the office into the field. A mobile GIS enables personnel to capture, store, update, manipulate, analyse, and display geographic information in the field. Compared to the usual way of collecting data in the field it offers many advantages:

- No paper map: Access to and visualization of digital data in flexible zoom level.
- No break between analogue and digital data.
- Providing field access to various GIS data. Access to data server with availability of mobile internet.
- Simplified inventory – possibility to compare between real object and existing GIS data.
- Decision making by using up-to-date, more accurate spatial data.

The following possibilities are enabled through positioning technologies:

- Query the GIS data based on actual location, overlay the data layer with actual position.
- Navigation: Field worker can determine their location and orientation in the field, using positioning sensor and georeferenced map or GIS data, both shown in the display.
- Surveying: Capture a feature by measuring a single position of a point feature, or specific geometry of a polyline or a polygon feature using the vertices, which describe the shape at the best
- Tracking: Geometrical data of a moved or driven route, with time stamp and other position sensor-based attributes.

Routing: Determine the “best” path (fastest, shortest), analyse the actual and the target location in relation to GIS topological and geometrical network data (route information network).

### **7.2 Hardware, components and technologies**

As possible IT platforms for mobile GIS applications generally all portable devices with an independent power supply are considered. In addition to devices in the consumer market (smartphones, tablet PCs, notebooks), the

spectrum ranges from PDAs to handheld devices that are specially designed for outdoor use, to rugged tablet PCs and notebooks.

Traditional mobile GIS applications fall into two basic categories dependent on the operating system and the kind of the mobile computer:

- Lightweight applications (function-reduced GIS, e.g. Esri ArcPad) running on handheld personal devices (TabletPC, Smartphone, PocktPC) with special operating system (OS) for mobile devices (e.g. Android, iOS and Windows mobile) TabletPC, and
- A ruggedised<sup>13</sup> Laptop with a powerful OS (e.g. Windows) that can also be operated via touchscreen, carried by a person or vehicle-mounted notebooks are used, allowing existing Windows-based GIS applications, such as Esri ArcGIS, to be used in the field.

Mobile GIS can integrate further modern technologies:

- Positioning sensor: Determine the position in respect to the digital data and maps, capture objects by surveying via GNSS or Total Station and navigation. To overcome GNSS reception restrictions or to enable indoor navigation, multi-sensor systems are also being developed or alternative beacon methods based on short-range radio (e.g., WLAN, Bluetooth, ultra-wideband) are used.
- Mobile internet: Full coverage of fast Internet (UMTS 3G network, LTE, or in future 5G), access to web-based GIS data resources (map server, feature server) and exchange of data with the enterprise or office.
- Barcode-scanner: Link geographic locations to information represented by the barcode.
- Digital camera: Link geographic positions to image information.

The demands on the hardware in outdoor use are very high. They often have to withstand environmental conditions such as cold, heat, dust, water and humidity for years to come. Commercially available everyday appliances from the consumer market are in most cases not designed for daily use under harsh environmental conditions. A long battery life and fall protection should also be considered for field service equipment.

The decision-making basis for a device can be, in addition to the size of a display, but also the operating technology and readability of the display:

- *Resistive touchscreens* can be operated with gloves. Any pen is usable as a precise input tool. This advantage can also be a disadvantage, as any contact of any objects is registered with the display. Classic zoom and rotation gestures, such as with smartphone use, are not implemented.
- *Capacitive touchscreens*, on the other hand, enable so-called multi-touch gestures with which, for example, elements can be rotated or scaled. The human finger is the primary input tool. In addition to the control by finger only conductive styli can be used. This is problematic because the touch input is not accurate enough, for example, to enter a vertex. In addition, external influences such as dust and rain make accessibility of capacitive touchscreens difficult.

Basically, it can be said that the larger the display, the higher the resolution should be, which, however, requires more computing power again.

Depending on the field of application, the criteria position accuracy, robustness, display, interfaces and operating system have different priorities. The more specific the requirements for these factors, the higher is the price in most cases. Therefore, it is necessary to create an exact requirement profile for the hardware.

### **7.3 Operating software**

Which platform is used depends mostly on the requirements of the software application. For example, on a Windows 10 tablet PC, the same desktop software may be used as on the desktop PC. Such a solution offers a high degree of convenience in terms of the display and saves training in new software. Handhelds, on the other hand, are smaller, lighter, and generally have longer battery life.

As an operating system for rugged GNSS handhelds, Windows Mobile is (still) the most prevalent. More powerful, rugged tablet PCs typically use desktop versions of Windows. Due to the proliferation of smartphones and tablet PCs in the private sector, operating systems such as iOS (Apple) and Android (Google) are pushing into the market for rugged handhelds.

Handheld devices, PDAs and Tablets based on Windows mobile OS, Google Android or Apple iOS are restricted by their size, particularly the display size, keyboard- and computing power (processor and RAM). Windows mobile OS and the devices themselves will not support Windows-based application for PC, such as Esri ArcGIS.

The Tablet PC digital pen and ink technology, touchscreen technology, text and voice recognition, and a Windows operating system for Laptops/PC enable Desktop-GIS as a high-end mobile computing solution for GIS.

---

<sup>13</sup> Protected against environmental conditions (e.g. dust, humidity, water, heat, cold) and resistant against mechanical influences (e.g. vibration, impact, shock, fall) according to different security levels (e. g. IP standard).

## 7.4 Mobile GIS software and geo-apps

In recent years, apps for mobile devices on Android or iOS are increasingly found on the market, enabling easy capturing of positions and descriptive attributes on devices that almost everyone owns anyway. Using the example of the market leader Esri, it can be shown that the focus is no longer only on a complex solution such as ArcPad, but in addition many Geo apps are marketed. ArcPad is complemented by many apps that splits its functionality has been split up into different geo-apps. For example, one for geometrical position detection, another for form entry of attributes, and so on.

For the functionality of the software, the rule "as little as possible, as much as necessary" applies. Due to the often small display sizes and the limited amount of available hardware resources, the focus is on clarity and ease of use for the mobile software. For occasional data collection in the field, the employees should not be overwhelmed by overloaded software.

Geo-Apps are applications that are available on smartphones. They are explicitly designed for the acquisition, processing and visualization of geodata. However, the "classical" division of tasks becomes more and more blurred, as in many applications the GNSS receiver is so integrated into the hardware and the application that it is no longer perceived as a single component. On the other hand, the synchronization (online or offline) of the data and the administration of user rights and roles (who can see or edit what data?) are becoming more and more important on mobile devices. There are offline solutions, whereby no web map service is used via a web browser, but an application on smartphones. This is necessary for working in areas with limited Internet connection. The integration of background maps and routing networks is often done in practice via APIs.

It has to be clarified for each application how the data exchange between mobile terminal and central database should take place. Which office software is used and by which means or in what time interval are the data used provided and updated? Windows for mobile based applications on light-weight computers, such as Esri ArcPad, are integrated into strategies to exchange field mapping datas with enterprise GIS (e.g. rules for synchronisation), such as Esri ArcGIS. The Esri Apps, on the other hand, are more tailored to the ArcGIS Online platform, making it easier to share data within an organization. For example, in the case of distributed data acquisition by several workers in the field, the data transferred back by them can be analyzed by the manager for overall work progress via an app.

## 7.5 Measurement-based GIS

Today's GIS almost all adopt the concept of coordinate GIS, which store only the coordinates of features. It assumes that it is possible to know the location exactly, but in practice exact location is not always known. Some GIS software developers have released GIS-extensions to integrate measurements in a GIS (ESRI ArcGIS SurveyAnalyst, Leica Mobile Matrix, TRIMBLE GPS Analyst). Such an extension consists of a set of tools for importing measurements and for defining the functional model between the measurements and the coordinates. These ideas led to measurement based systems, on which various researchers are still working (Hintz et al., 1996, Goodchild, 1999, Goodchild, 2004).

A measurement-based GIS requires storing the original measurements instead of derived data such as coordinates. The data structure of a geodatabase is extended to store survey measurements, survey computation, resulting survey points and rules to attach the survey point with a GIS feature. Multiple coordinates (with quality information) can be computed for one survey point displayed at the position of the mean or a user defined coordinate. Measuring features produces a survey point that is coincident with a feature or one of its vertices (Leica 2007). It is possible to adjust measurements and to attach boundaries to points. Because the GIS objects are coupled to the measurement-based survey points through topological rules in the geodatabase, it is also possible to add new measurements and re-adjust the surrounding points. After a new adjustment, all features in the GIS based on this survey network can be updated. The system stores information about the accuracy of the coordinates with each boundary point and thus allows the computation of accuracy for derived spatial features such as areas or distances (Navratil, 2004).

# 8 Digitising and georeferencing

## 8.1 Scanning

*"...a scanner is a device that analyzes images, printed text, or handwriting, or an object (such as an ornament) and converts it to a digital image."<sup>14</sup>*

<sup>14</sup> [http://en.wikipedia.org/wiki/Image\\_scanner](http://en.wikipedia.org/wiki/Image_scanner)

After placing a map on a glass plate, a light beam passes over it and measures the reflected light intensity. The result is a matrix of pixels. This matrix consists of  $n$  columns and  $k$  rows. The counting of the columns starts usually in the upper or lower left corner. Important criteria for scanning a map are:

- Resolution
- Colour depth
- File size
- File format and compression

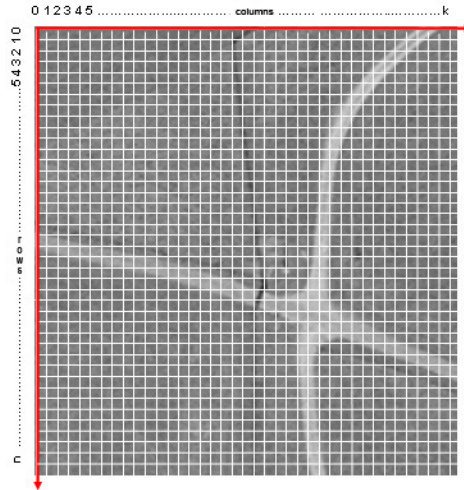


Figure 48: Image- or raster matrix

### 8.1.1 Resolution

Usually resolution is expressed with the unit “dots per inch” (dpi), e.g., based on 1 inch (2,54cm) 100dpi correspond to 39 Pixels per cm or 4 Pixels per mm. The resolution of the image depends on the planned usage, e.g., for reproduction 300dpi is sufficient, and on the kind of the sample which has to be digitised, e.g., legibility of characters. Figure 49 shows an example of different resolutions.



72dpi



100dpi



300dpi

Figure 49: Different resolutions

### 8.1.2 Colour depth and file size

The colour depth is a computer graphics term describing the number of bits used to represent the colour of a single pixel in a bitmapped image. Thereby a higher colour depth gives a broader range of distinct colours, but also a larger file size (Figure 50 and Table 5).

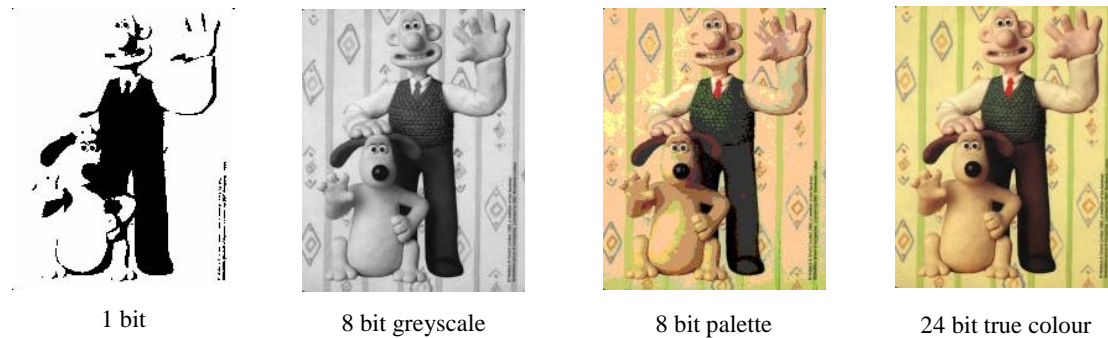


Figure 50: Different colour depths

Colour depth	Viewable colours	Name	Filesize of a map of 10×10cm at 300dpi
1 bit	$2^1 = 2$	Monochrome (s/w)	ca. 170 Kb
8 bit	$2^8 = 256$	Greyscale/Palette	ca. 1360 Kb
24 bit	$2^{24} = 16.777.216$	True colour	ca. 4079 Kb

Table 5: Colour depth and file size

### 8.1.3 File formats and compression

Using file formats without a publicly available specification can be costly. Any file format used should therefore have an open definition. Furthermore, the format must be widespread and to use this data in a GIS, additional information such as calibration data, metadata and geoinformation should have been integrated. It must be error-tolerant and it must be compatible with modern and lossless compression algorithms.

Image compression is the application of data compression on digital images. In effect, the objective is to reduce redundancy of the image data in order to be able to store or transmit data in an efficient form.

Image compression can be lossy or lossless. Lossless compression is sometimes preferred for artificial images such as technical drawings, etc. This is because lossy compression methods, especially when used at low bit rates, introduce compression artefacts. Lossless compression methods may also be preferred for high value content, such as medical imagery or image scans made for archival purposes. Lossy methods are especially suitable for natural images such as photos in applications where minor (sometimes imperceptible) loss of fidelity is acceptable to achieve a substantial reduction in bit rate.

The most commonly used formats are:

- JPEG – Joint Photographic Experts Group (\*.JPG files)
- PNG – Portable Network Graphics (\*.PNG files)
- GIF – Graphic Interchange Format (\*.GIF files)
- TIFF/GeoTIFF – Tag Image File Format (\*.TIF, \*.TFW files)
- IMG – ERDAS format (\*.IMG)

## 8.2 Georeferencing

The most frequent use of plane coordinate transformation techniques takes place during the process of converting digitised cartographic data – tablet digitised or scanned maps – into georeferenced Cartesian coordinates. Tablet-traced vector features or scanned raster data have imposed upon them an arbitrary Cartesian coordinate system. The coordinates of tablet-digitised data are often in inches, with the origin of the coordinate system in the lower-left of the tablet. Scanned digital data are made up of square cells, or pixels, of colour (or shades of grey). The number of cells per inch is controlled during the scanning process. The number of rows and columns of cells defines a Cartesian space, usually with the origin in the upper left or the lower left corner.

To transform this newly digitised data (either traced or scanned) the most common method is known as an **affine transformation**. It is based on two linear or first order polynomial equations with six coefficients. It allows for the differential scaling, skewing, rotation, and translation of the data.



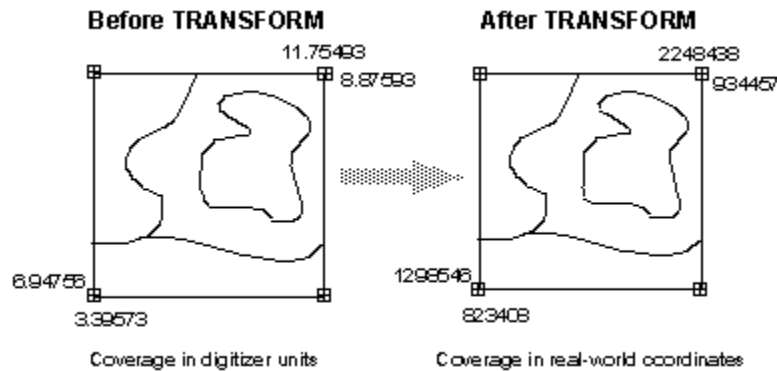


Figure 51: Converting data in digitiser units to georeferenced Cartesian, or real-world, coordinates by geometric transformation (source: ESRI, 2007)

A simpler linear transformation method is the **similarity, or conformal, method**. It is not usually applied to raster data. A more complex linear transformation is the projective method. It is only applicable for high-altitude aerial photography or aerial photos of flat terrain.

If the transformation requires a curving of the data, then a **second- or third-order polynomial equation** is used. This is often referred to as **warping**, usually in the context of georeferencing raster images. It is recommended that second-order and higher polynomial transformations be used with caution. They can significantly deform all but the control point locations of the data.

In order to implement a transformation method a set of control points is determined. These are pairs of coordinates that are known in both the digitised coordinate system and in the georeferenced Cartesian coordinate system being transformed into. For example, the XY digitising tablet coordinates for road intersections can be determined during the digitising process. The corresponding georeferenced Cartesian coordinates can be measured in the field with a GNSS receiver. Each transformation method requires a minimum number of control point pairs in order for the transformation equation coefficients to be calculated. The equations are then formulated that will convert the digitised values of the control points to the georeferenced values of the control points. These equations are then used to convert the remaining coordinates that define all of the digitised data into georeferenced coordinates; scaling, rotating, and shifting all features as necessary. Straight lines remain straight after a linear transformation. The number and location of control points is critical. Their placement determines the type of transformation possible and the accuracy of the result.

A few rules of thumb when selecting control points:

- use control points as far apart from each other as possible, on the outer edge of the map, and, if more than two points are used, not in a single line,
- use easily recognizable points in the image or data set that is being transformed and the original source data,
- use source data recorded in an appropriate projection to make their georeferenced location clearly identifiable

An affine transformation can differentially scale the data, skew it, rotate it, and translate it. The graphic below illustrates the four possible changes (Figure 52).

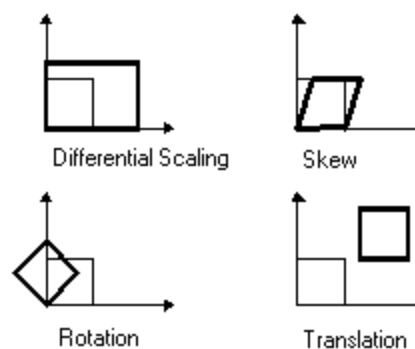


Figure 52: Four different transformation parameters

The affine transformation function is:

$$\begin{aligned}x' &= Ax + By + C \\y' &= Dx + Ey + F\end{aligned}$$

where  $x$  and  $y$  are coordinates of the input layer and  $x'$  and  $y'$  are the transformed coordinates.  $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$  and  $F$  are determined by comparing the location of source and destination control points. They scale, skew, rotate, and translate the layer coordinates. The affine transformation requires a minimum of three control points.

The similarity transformation scales, rotates, and translates the data. It will not independently scale the axes, nor will it introduce any skew. It maintains the aspect ratio of the features transformed.

The similarity transform function is:

$$\begin{aligned}x' &= Ax + By + C \\y' &= -Bx + Ay + F\end{aligned}$$

where:

$$A = s * \cos t \quad B = s * \sin t \quad C = \text{translation in x direction} \quad F = \text{translation in y direction}$$

and

$s$  = scale change (same in  $x$  and  $y$  directions)

$t$  = rotation angle, measured counter-clockwise from the  $x$ -axis.

A similarity transformation requires a minimum of two control points.

The projective transformation is based upon a more complex formula that requires a minimum of four control points.

$$\begin{aligned}x' &= (Ax + By + C) / (Gx + Hy + 1) \\y' &= (Dx + Ey + F) / (Gx + Hy + 1)\end{aligned}$$

### 8.2.1 Residual and root mean square (RMS)

The transformation parameters are a best fit between the source and destination control points. If you use the transformation parameters to transform the actual source control points, the transformed output locations won't match the true output control point locations. This is called the residual error; it is a measure of the fit between the true locations and the transformed locations of the output control points. This error is generated for each displacement link.

A **root mean square error RMS** is calculated for each transformation performed, which indicates the quality of the derived transformation. The following example illustrates the relative location of four destination control points and the transformed source control points (Figure 53):

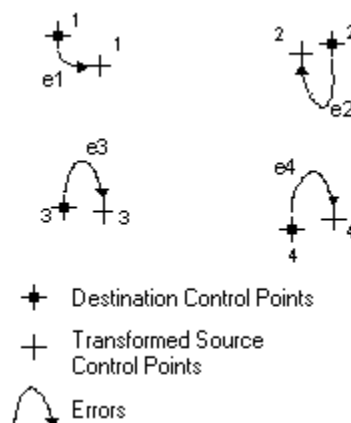


Figure 53: Errors in transformations

The RMS error measures the errors between the destination control points and the transformed locations of the source control points:

$$\text{RMS error} = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2}{n}}$$

The transformation is derived using least squares, so more links can be given than are necessary.

### 8.3 Digitising

Digitising is the method of converting information from one format to another. Traditionally, digitising has meant the creation of a spatial dataset from a hardcopy source such as a paper map or a plan. On-screen digitising is the creation of a spatial dataset by tracing over features displayed on a computer monitor using a mouse. In both cases, the newly created dataset picks up the spatial reference of the source document. Existing maps can be digitised using a scanner or tablet digitiser. Raster data are obtained from a scanner while vector data are produced using a digitiser (Figure 54).

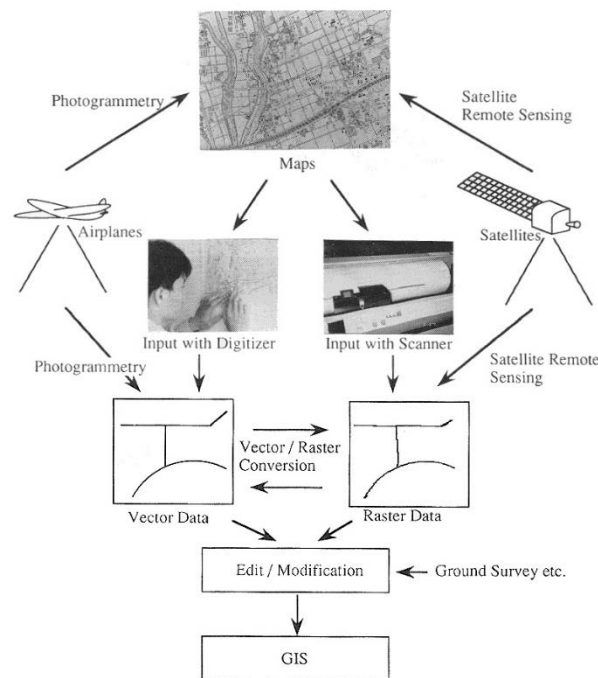


Figure 54: A process of input / update of geographic data <sup>15</sup>

### References

- Burt, WA, Kaplan, D. J., Keenly, R. R., Reeves, J. F., Shaffer, F. B. (1965): Mathematical Considerations Pertaining to the Accuracy of Position Location and Navigation Systems, Research Memorandum, Stanford Research Institute, November 1965. The essential content of the report is published in: Bowditch: American Practical Navigator, Vol.1 1977.
- DMA (1991): DMA Technical Report 8400.1, Error theory as applied to Mapping, Charting and Geodesy, Defense Mapping Agency (DMA), May 1991, URL: [gis-lab.info/docs/dma-tr-tr8400\\_1.pdf](http://gis-lab.info/docs/dma-tr-tr8400_1.pdf), URL visited: 05.08.2019.
- Encyclopaedia Astronautica (2019): BeiDou, URL: <http://www.astronautix.com/b/beidou3.html>, (visited: 03.08.2019).
- EPSG (2019): EPSG Geodetic Parameter Dataset, Version 9.6.3, URL: <https://www.epsg-registry.org/> (visited: 12.06.2019)
- ESA (2019): ESA Navipedia, URL: [https://gssc.esa.int/navipedia/index.php/Main\\_Page](https://gssc.esa.int/navipedia/index.php/Main_Page). Link visited: 25.07.2019
- ESRI (2004): ArcGIS® 9, Understanding Map Projections, ESRI Library (ArcGIS 9 Documentation).
- ESRI (2007): ArcGIS 9.2 Desktop Help, URL: <http://webhelp.esri.com/arcgisdesktop/> (visited: 28.08.2007).
- Goodchild, M. F. (1999): Measurement-Based GIS, International Symposium on Spatial Data Quality, Hong Kong, Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University.
- Goodchild, M. F. (2004): A general framework for error analysis in measurement-based GIS. Journal of Geographical Systems, Volume 6, Number 4 / December, 2004, Springer, Berlin – Heidelberg.
- GSA (2018): Galileo Services, European Global Navigation Satellite System Agency, URL: <https://www.gsa.europa.eu/galileo/services>. Page updated: Sep 25, 2018. URL visited: 26.07.2019.

<sup>15</sup> Source: <http://www.africanconservation.org/dcforum/DCForumID5/310.html>

- Harre, I. (1987): Modelling of Navigation Errors – Development of a Circular Error Algorithm Session of 'DGON-Schiffahrtskommission', Hamburg, 01.04.87, published in 'Ortung und Navigation', No. III, 1987, Düsseldorf
- Harre, I. (2001): A Standardized Algorithm for the Determination of Position Errors by the Example of GPS with and without 'Selective Availability' in The International Hydrographic Journal, Vol. 2, No. 1 (New Series), June 2001.
- Harter, L.P. (1960): Circular Error Probabilities; Journal American Statistical Association, Vol. 55, December 1960.
- Hintz, R. J., Wahl, J. L., Wurm, K., McKay, D. (1996): Geographic Measurement Management: An Operational Measurement-Based Land Information System. ASPRS/ACSM Annual Convention, Baltimore, MD, 1996. URL: <http://www.cadastral.com/gmm96ah.htm> (visited: 03.09.2007).
- Huisman, L. (2004): Integrating network adjustment software within ArcGIS Survey Analyst - Geo-DBMS Case Study. URL:<http://www.gdmc.nl/publications/index.html> (visited: 29.08.2007).
- Ihde, J., Luthardt, J., Boucer, C., Dunkley, P., Farrell, B., Gubler, E., Torres, J.: European Spatial Reference Systems - Frames for Geoinformation Systems. URL: <http://www.crs-geo.eu/pub01EuropeanSpatialRefernceSystems.pdf> (visited: 26.06.2019).
- IOGP (2012): Surveying and Positioning Guidance Note Number 7, part 1: Using the EPSG Geodetic Parameter Dataset, Version 8, August 2012. Association of Oil & Gas Producers Surveying and Positioning Committee. OGP Publication 373-7-1, 2012. URL: <http://www.epsg.org/Guidancenotes.aspx> (visited: 03.06.2019).
- IOGP (2019): Guidance Note Number 7, part 2: Coordinate Conversions and Transformations including Formulas, Revised February 2019. Association of Oil & Gas Producers Surveying and Positioning Committee. OGP Publication 373-7-2, 2019. URL: <http://www.epsg.org/Guidancenotes.aspx> (visited: 03.06.2019).
- ISO (International Organization for Standardisation) (2003): ISO 19115:2007, Geographic information -- Metadata.
- ISO (International Organization for Standardisation) (2004): ISO 19116:2004, Geographic information -- Positioning services.
- ISO (International Organization for Standardisation) (2005): ISO19127:2005, Geographic information -- Geodetic codes and parameters.
- ISO (International Organization for Standardisation) (2019): Geographic Information - Spatial referencing by coordinates [ISO 19111:2019].
- Leica Geosystems (2007): Leica Mobile Matrix – Mobile Data Collection and Maintenance Solution (Technical White Paper), Leica Geosystems AG, 2007.
- Mahammad, Sk. Sazid, Ramakrishnan, R.: GeoTIFF – A standard image file format for GIS applications. Geospatial World, 2009. URL: <https://www.geospatialworld.net/article/geotiff-a-standard-image-file-format-for-gis-applications/> (visited: 03.06.2019).
- Navratil, G., Franz, M., Pontikakis, E. (2004): Measurement-Based GIS Revisited, 7th AGILE Conference on Geographic Information Science” 29 April-1May 2004, Heraklion, Greece. Poster Session URL: [http://www.geoinfo.tuwien.ac.at/publications/index.php?by\\_author:Navratil%2C\\_Gerhard](http://www.geoinfo.tuwien.ac.at/publications/index.php?by_author:Navratil%2C_Gerhard) (visited: 29.08.2007).
- OGC (Open Geospatial Consortium Inc.) (2010): The OGC Abstract Specification – Topic2: Spatial referencing by coordinates, Version 4.0.0, 27.04.2010. URL: <https://docs.opengeospatial.org/as/18-005r4/18-005r4.html> (visited: 12.06.2019).
- Resnik, B., Bill, R. (2018): Vermessungskunde für den Planungs-, Bau- und Umweltbereich, 4nd Edition, Wichmann, Heidelberg.
- Sanz Subirana, J., Juan Zornoza, J.M. and Hernández-Pajares, M.: GNSS DATA PROCESSING, Volume I: Fundamentals and Algorithms, European Space Agency, TM-23/1, 2013.
- Teunissen, P. J.G., Montenbruck, O. (2017): Springer Handbook of Global Navigation Satellite Systems, Springer, 2017.
- Torge, W. (1991): Geodesy, 2nd Edition, de Gruyter, Berlin – New York.
- Vaníček, P., Krakiwsky, E.J. (1986). Geodesy: the concepts, 2nd edition. North Holland, Amsterdam.
- Wikipedia (2007): Wikipedia - the free encyclopedia, URL: <http://en.wikipedia.org/wiki/> (visited: 28.08.2007).



**Part C**

**Remote Sensing**

Dr.-Ing. Görres Grenzdörffer



# 1 Introduction

This chapter on remote sensing does not fully cover all aspects of remote sensing, but it gives details and background information on selected important issues. The text and the figures are derived from various sources. Most important to mention are the text books from Jensen (2005, 2007).

## 1.1 Why remote sensing?

There are several good reasons to use remote sensing in several disciplines for a large variety of applications. With remote sensing, you may:

- Observe: land cover, boundaries, threats, damage, topography, ...
- Measure: areas, distances, height/elevation, ...
- Detect: fires, resource use violations, ...
- Monitor: changes in forest cover, crop and range condition, land use, ...
- Classify: into vegetation and land use categories, habitats, ...

## 1.2 What is remote sensing?

### 1.2.1 Definitions

There are many possible definitions of what 'Remote Sensing' actually is, as the following quotes demonstrate:

F.F. Sabins in his book "Remote sensing: principles and interpretation" (1996) defines it as follows:

*"Remote Sensing is the science of acquiring, processing and interpreting images that record the interaction between electromagnetic energy and matter."*

Lillesand et al in their book "Remote Sensing and Image Interpretation" (2005) even define it as an art:

*"Remote Sensing is the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation."*

Remote sensing may also be defined as a process or workflow from data capture to data analysis and data presentation, see Figure 55.

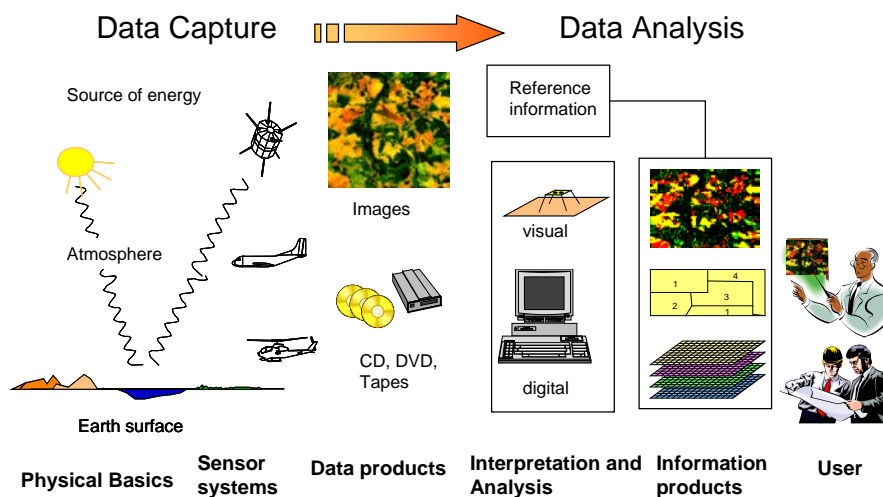


Figure 55: The remote sensing workflow

Beside the spaceborne or airborne part of remote sensing, ground work is very important for remote sensing applications and science. Thereby:

- *In situ* and collateral data necessary to calibrate the remote sensing data and/or judge its geometric, radiometric, and thematic characteristics are collected.
- Remote sensing data are collected passively or actively using analogue or digital remote sensing instruments, ideally at the same time as the *in situ* data.



### 1.3 History of remote sensing

The history of remote sensing is closely tied with the development of photography (dating back to the 19<sup>th</sup> century), analogue and digital image acquisition, and, importantly, with digital aerial and spaceborne technology. Earth observation with satellites started in the Cold War, primarily with satellite photographs. The following list shows some important events in the development of remote sensing:

- 1839 – Invention of photography (Niépce, Daguerre, Arago)
- 1901 – Stereo photogrammetry, first device for stereo measurement (Stereo comparator by Pulfrich)
- 1955 – Production of first orthophotos from aerial imagery (Bean)
- 1972 – First Landsat satellite launched
- 1986 – SPOT satellite launched
- 1988 – Indian Remote Sensing Satellite launched
- 1995 – Radarsat launched
- 1999 – IKONOS satellite launched and NASA launched Terra satellite
- 2006 – TerraSAR-X and Radarsat 2 launched
- 2010 – Unmanned aerial systems (UAS)
- 2015 – Sentinel 2

Today many new satellite instruments (10+ per year), are being developed and launched under governmental supervision or by private enterprises. The general trend is towards ever higher spatial resolution, not only in the visible part of the spectrum but also with radar satellites. With UAS a new and very flexible method of close range remote sensing is now available.

## 2 Physical basics of remote sensing

### 2.1 Light, atmosphere and reflection properties of the Earth surface

Energy recorded by remote sensing systems undergoes fundamental *interactions* that must be understood in order to properly pre-process and interpret remotely sensed data. For example, if the energy being remotely sensed comes from the Sun, the energy

1. is radiated by atomic particles at the source (the Sun),
2. travels through the vacuum of space at the speed of light,
3. interacts with the Earth's atmosphere,
4. interacts with the Earth's surface,
5. interacts with the Earth's atmosphere once again,
6. finally reaches the remote sensor, where it interacts with various optics, filters, film emulsions, or detectors.

The Sun approximates a 6000K blackbody with a dominant wavelength of 0.48 $\mu\text{m}$  (green light). The Earth approximates a 300K blackbody with a dominant wavelength of 9.66 $\mu\text{m}$ . The 6000K Sun produces 41% of its energy in the visible region from 0.4 - 0.7 $\mu\text{m}$  (blue, green, and red light). The other 59% of the energy is in wavelengths shorter than blue light (<0.4 $\mu\text{m}$ ) and longer than red light (>0.7 $\mu\text{m}$ ). Eyes are only sensitive to light from the length of 0.4 to 0.7 $\mu\text{m}$ . Remote sensor detectors can be made sensitive to energy in the non-visible regions of the spectrum.

Some types of electromagnetic radiation easily pass through the atmosphere, while other types do not. The ability of the atmosphere to allow radiation to pass through it, is referred to as its *transmissivity*, and varies with the wavelength/type of the radiation. The gases that comprise our atmosphere absorb radiation in certain wavelengths while allowing radiation with differing wavelengths to pass through.

The areas of the EM spectrum that are absorbed by atmospheric gases such as water vapour, carbon dioxide, and ozone are known as absorption bands. In Figure 57, *absorption bands* are represented by a low transmission value that is associated with a specific range of wavelengths.

In contrast to the absorption bands, there are areas of the electromagnetic spectrum where the atmosphere is transparent (little or no absorption of radiation) to specific wavelengths. These wavelength bands are known as "**atmospheric windows**" since they allow the radiation to easily pass through the atmosphere to the Earth's surface, (Figure 57). The combined effects of atmospheric absorption, scattering, and reflectance (Figure 58) reduce the amount of solar irradiance reaching the Earth's surface at sea level.

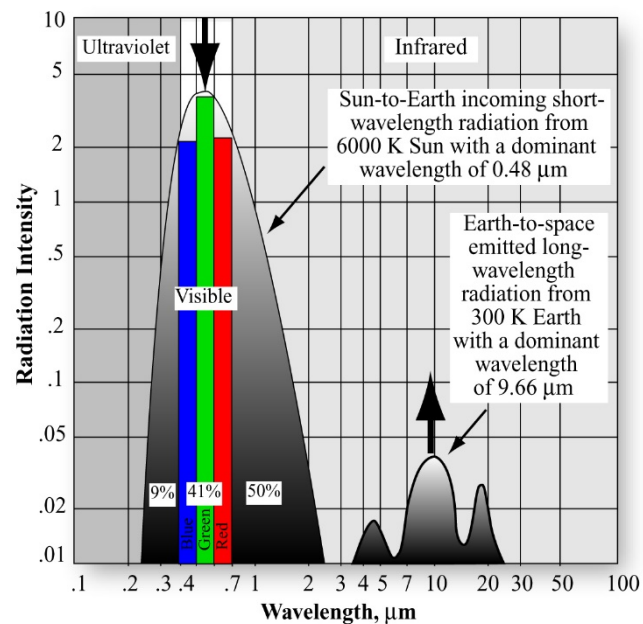


Figure 56: Radiant intensity of the sun (Jensen, 2007)

Most remote sensing instruments on aircraft or space-based platforms operate in one or more of these windows by making their measurements with detectors tuned to specific frequencies (wavelengths) that pass through the atmosphere. When a remote sensing instrument has a line-of-sight with an object that is reflecting sunlight or emitting heat, the instrument collects and records the radiant energy. While most remote sensing systems are designed to collect reflected radiation, some sensors, especially those on meteorological satellites, directly measure absorption phenomena, such as those associated with carbon dioxide (CO<sub>2</sub>) and other gases. The atmosphere is nearly opaque to EM radiation in part of the mid-IR and all of the far-IR regions. In the microwave region, by contrast, most of this radiation passes through unimpeded, so radar waves reach the surface (although weather radars are able to detect clouds and precipitation because they are tuned to observe backscattered radiation from liquid and ice particles).

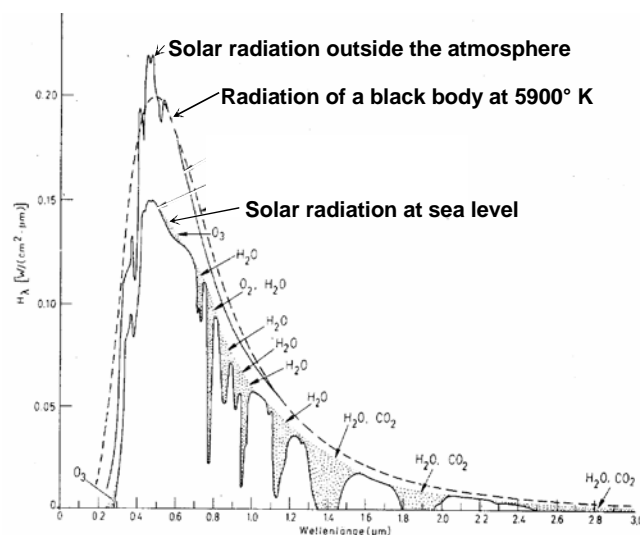


Figure 57: Atmospheric windows - absorption bands of the Earth atmosphere

There is always an **angle of incidence** associated with the incoming energy that illuminates the terrain and an **angle of reflection** from the terrain to the sensor system. This **bidirectional** nature of remote sensing data collection is known to influence the spectral and polarization characteristics of the at-sensor radiance,  $L$ , recorded by the remote sensing system.

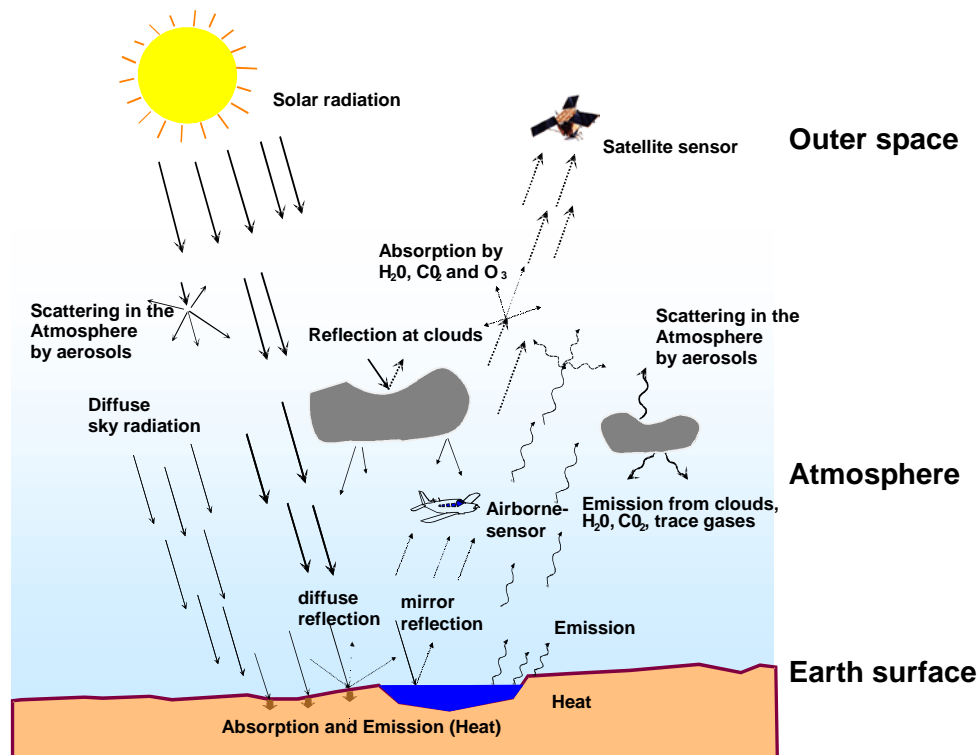


Figure 58: Reflection, scattering and absorption processes in the atmosphere and at the Earth surface

### 2.1.1 Atmospheric correction

**Radiometric correction** requires knowledge of electromagnetic radiation principles and what interactions take place during the remote sensing data collection process. To be exact, it also involves knowledge about the terrain *slope* and *aspect* and *bi-directional reflectance characteristics* of the scene. There are several ways to atmospherically correct remotely sensed data. Some are relatively straightforward while others are complex, being founded on physical principles and requiring a significant amount of information to function properly. This discussion will focus on two major types of atmospheric correction:

- *Absolute atmospheric correction*
- *Relative atmospheric correction*

There are various methods that can be used to achieve absolute or relative atmospheric correction. The following sections identify the logic, algorithms, and problems associated with each methodology.

Solar radiation is largely unaffected as it travels through the vacuum of space. When it interacts with the Earth's atmosphere, however, it is selectively scattered and absorbed. The sum of these two forms of energy loss is called *atmospheric attenuation*. Atmospheric attenuation may 1) make it difficult to relate hand-held *in situ* spectroradiometer measurements with remote measurements, 2) make it difficult to extend spectral signatures through space and time, and 3) have an impact on classification accuracy within a scene if atmospheric attenuation varies significantly throughout the image.

The general goal of *absolute radiometric correction* is to turn the digital brightness values recorded by a remote sensing system into *scaled surface reflectance* values. These values can then be compared or used in conjunction with scaled surface reflectance values obtained anywhere else on the planet.

Much research has been carried out to address the problem of correcting images for atmospheric effects. These efforts have resulted in a number of *atmospheric radiative transfer codes (models)* that can provide realistic estimates of the effects of atmospheric scattering and absorption on satellite imagery. Once these effects have been identified for a specific date of imagery, each band and/or pixel in the scene can be adjusted to remove the effects of scattering and/or absorption. The image is then considered to be *atmospherically corrected* (Figure 59).

Unfortunately, the application of these codes to a specific scene and date also requires knowledge of both the sensor spectral profile and the atmospheric properties at the same time. Atmospheric properties are difficult to acquire even when planned. For most historic satellite data, they are not available. Even today, accurate scaled surface reflectance retrieval is not operational for the majority of satellite image sources used for land-cover change detection. An exception is NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), for which surface reflectance products are available. Nevertheless, we will proceed with a general discussion of the important issues associated with absolute atmospheric correction and then provide examples of how absolute radiometric correction is performed.

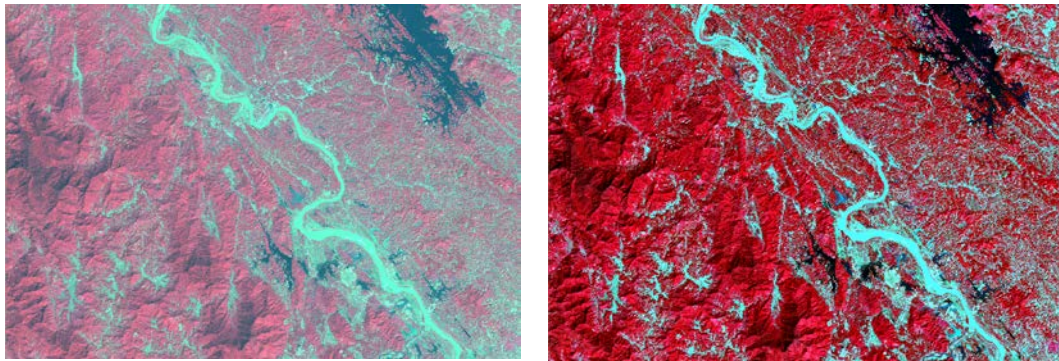


Figure 59: Atmospheric correction with ATCOR of Landsat 7 ETM scene WRS 127/045, 11.04.2000

Most current *radiative transfer-based atmospheric correction algorithms* can compute much of the required information if a) the user provides fundamental atmospheric characteristic information to the program or b) certain atmospheric absorption bands are present in the remote sensing dataset. For example, most radiative transfer-based atmospheric correction algorithms require that the user provide:

- latitude and longitude of the remotely sensed image scene,
- date and exact time of remote sensing data collection,
- image acquisition altitude (e.g. 20km above ground level, AGL),
- mean elevation of the scene (e.g., 200m above sea level, ASL),
- an atmospheric model (e.g. mid-latitude summer, mid-latitude winter, tropical),
- radiometrically calibrated image radiance data (i.e. data *must* be in the form  $Wm^2mm^{-1}sr^{-1}$ ),
- data about each specific band (i.e. the mean and full-width at half-maximum (FWHM)), and
- local atmospheric visibility at the time of remote sensing data collection (e.g. 10km, obtained from a nearby airport if possible).

These parameters are then input to the atmospheric model selected (e.g. mid-latitude summer) and used to compute the absorption and scattering characteristics of the atmosphere at the instance of remote sensing data collection. These atmospheric characteristics are then used to invert the remote sensing radiance to *scaled surface reflectance*. Many of these atmospheric correction programs derive the scattering and absorption information they require from robust atmosphere radiative transfer code such as MODTRAN 4+ or Second Simulation of the Satellite Signal in the Solar Spectrum (6S).

### 2.1.2 Remote sensor resolution

*Resolution* is a very important term in remote sensing, and it has several meanings attached to it:

- **Spatial Resolution or ground resolution** - the size of the pixel in the field-of-view, e.g.  $10 \times 10m$ .
- **Spectral Resolution** - the number and size of spectral regions the sensor records data in, e.g. blue, green, red, near-infrared thermal infrared, microwave (radar).
- **Radiometric Resolution** - the sensitivity of detectors to small differences in electromagnetic energy.
- **Temporal Resolution** - how often the sensor acquires data, e.g. every 16 days.

For any kind of remote sensing project, remote sensing data with appropriate resolution has to be selected. In one case this may mean image data with the highest spatial resolution possible, in other cases a high temporal resolution is important.

## 2.2 Spectral signatures

For any given material, the amount of solar radiation that it reflects, absorbs, transmits, or emits varies with wavelength. When that amount (usually intensity, as a percent of maximum) coming from the material is plotted over a range of wavelengths, the connected points produce a curve called the material's *spectral signature*. The spectral signature is used for the separation of different materials by the analysis of multispectral or hyperspectral data.

Vegetation provides a unique spectral signature due to the reflectance properties of the leaves. Leaf organs are partially transparent. In the visible part of the spectrum the light is absorbed by the chlorophyll (blue, red). Due to this fact human see vegetation as green. In the infrared part of the spectrum boundary reflection (cell boundaries, hollows) results in high reflection values. Figure 60 shows the different factors that may influence the shape of the reflectance curve of healthy and unhealthy leaves.

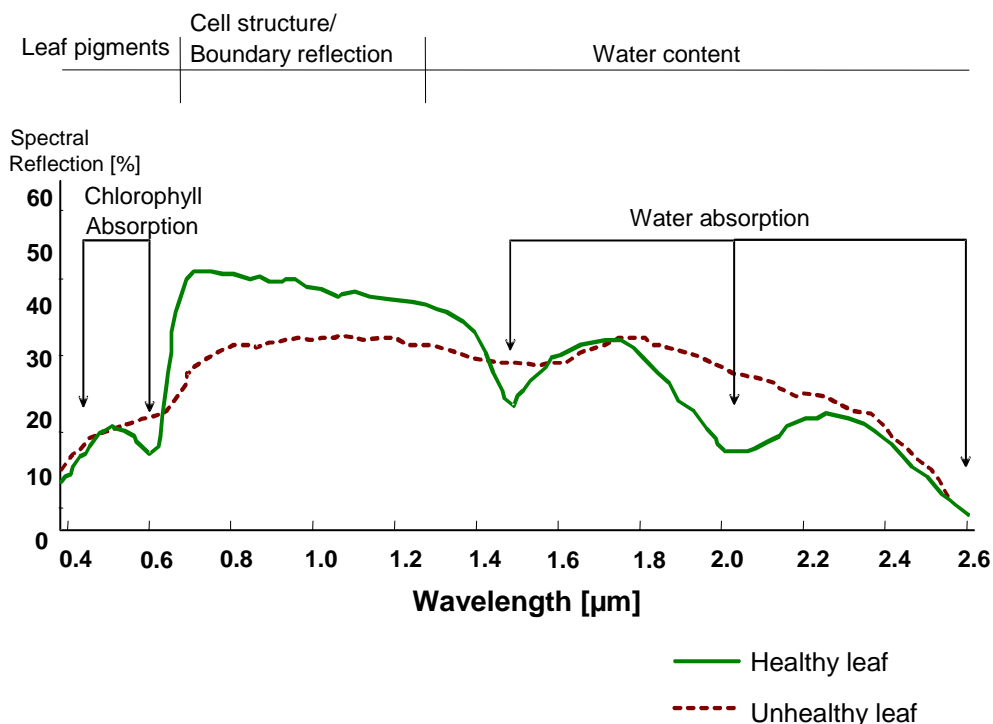


Figure 60: Spectral signature of a leaf

Due to these facts we may use remote sensing to determine the vitality and the healthiness of vegetation. While the general patterns may or may not remain the same, the spectra of features change over time due to the phenology of the vegetation.

### 2.2.1 Indices

However, (depending on the sensor type) we can recognize spectral patterns for vegetated versus non-vegetated areas, for certain classes of vegetation, and geologic and mineral properties. Thereby indices between two or more spectral bands are commonly used. The first vegetation index developed was the **simple ratio (SR)**. The SR computes the Infrared/Red ratio. It takes advantage of the inverse relationship between chlorophyll absorption of red radiant energy and increased reflectance of near-infrared energy for healthy plant canopies.

$$SR = \frac{NIR}{red}$$

The **Normalized Difference Vegetation Index (NDVI)** is the most common vegetation index.

$$NDVI = \frac{NIR - red}{NIR + red}$$

It can be seen from its mathematical definition that the NDVI of an area containing a dense vegetation canopy will tend to positive values (say 0.3 to 0.8) while clouds and snow fields will be characterized by negative values of this index. Other targets on Earth visible from space include free standing water (e.g., oceans, seas, lakes and rivers) which have a rather low reflectance in both spectral bands (at least away from shores) and thus result in very low positive or even slightly negative NDVI values, soils which generally exhibit a near-infrared spectral reflectance somewhat larger than the red, and thus tend to also generate rather small positive NDVI values (say 0.1 to 0.2). In addition to the simplicity of the algorithm and its capacity to broadly distinguish vegetated areas from other surface types, the NDVI also has the advantage of compressing the size of the data to be manipulated by a factor 2 (or more), since it replaces the two spectral bands by a single new field (often encoded using 8 bits instead of the 10 or more bits of the original data).

Using the NDVI for quantitative assessments (as opposed to qualitative surveys as indicated above) raises a number of issues that may seriously limit the actual usefulness of this index if they are not properly addressed. Also, the NDVI has tended to be over-used (if not abused) in applications for which it was never designed.

Recent emphasis has been given to the development of improved vegetation indices that may take advantage of calibrated hyperspectral sensor systems such as the moderate resolution imaging spectrometer MODIS. The improved indices incorporate a *soil adjustment factor* and/or a *blue band for atmospheric normalization*. The **soil adjusted vegetation index (SAVI)** introduces a soil calibration factor,  $L$ , to the NDVI equation to minimize soil background influences resulting from first order soil-plant spectral interactions (Huete, 1988).

$$SAVI = \frac{(1+L)(NIR - red)}{NIR + red + L}$$

An  $L$  value of 0.5 minimizes soil brightness variations and eliminates the need for additional calibration for different soils (Huete et al., 1994). Many indices may be found in the literature for a variety of applications. E.g. Rock et al. (1986) utilized a **Moisture Stress Index (MSI)** based on the Landsat Thematic Mapper near-infrared and middle-infrared bands

$$MSI = \frac{MidIR_{TM5}}{NIR_{TM4}}$$

### 3 Sensor systems

#### 3.1 Earth observation satellite systems

Nowadays some 80+ Earth observation satellites with different sensors and resolutions continuously acquire images from the Earth. Satellites move in an elliptical path with the planet at one of the foci of the ellipse. Most Earth observation satellites fly in circular orbits, because if images of different locations are to be suitable for comparison, they must be acquired from the same altitude. The orbit must therefore be circular, or have a constant altitude relative to the Earth's surface. The most important types are satellites in an either polar (or near-polar) orbit (= Sun-synchronous) or in a geostationary orbit above the equator. Because the valid comparison of images of a given location acquired on different dates depends on the similarity of the illumination conditions, the orbital plane must also form a constant angle relative to the sun direction. This is achieved by ensuring that the satellite overflies any given point at the same local time, which in turn requires that the orbit be sun-synchronous (e.g. descending node at 10:30 a.m. for SPOT satellites). The basic characteristics of satellites operating in the optical spectrum are related to the following properties:

- Spatial resolution
- Coverage area / swath width
- Spectral resolution
- Radiometric properties
- Pointing
- Revisit period
- Data availability

While the trend in remote sensing satellites is towards higher ground resolution, the temporal resolution is getting lower. Therefore high resolution satellites are all equipped with pointing devices enabling oblique images along and across the orbit.

Because the Landsat satellite was the first Earth observation satellite, which is still widely used especially for change detection purposes, some necessary details are given in the next paragraph.

##### 3.1.1 Landsat

The Landsat programme is the longest running enterprise for acquisition of imagery of the Earth from space. The first Landsat satellite was launched in 1972; the most recent, Landsat 8, was launched on 15<sup>th</sup> April 2013. The instruments on the Landsat satellites have acquired millions of images.<sup>16</sup> The Landsat program will be continued with the launch of a new satellite, planned for December 2020. The instrument characteristics of Landsat 9 will be the same as the current Landsat 8. The almost 50-year record of images provides a unique resource for people who work in agriculture, geology, forestry, regional planning, education, mapping, and global change research.

Landsats 1, 2, and 3 orbited at an altitude of 920km. These satellites circled the Earth every 103 minutes, completing 14 orbits a day. It took 18 days to provide nearly complete coverage of the Earth's surface with 185-km image swaths. The primary sensor aboard Landsats 1, 2, and 3 was the Multispectral Scanner (MSS). The resolution of the MSS sensor was approximately 80 meters, with four bands of spectral coverage ranging from the visible green to the near-infrared (IR) wavelengths.

The **Landsat Thematic Mapper (TM)** is a sensor carried onboard Landsat 4 and 5 and has acquired images of the Earth nearly continuously from July 1982 to the present, with a 16-day repeat cycle. Landsat TM image data consists of seven spectral bands with a spatial resolution of 30 meters for bands 1 to 5 and band 7. Spatial resolution for band 6 (thermal infrared) is 120 meters, but band 6 data are oversampled to 30 meter pixel size. Approximate scene size is 170 km north-south by 183 km east-west. All Landsat 5 / 8 data are referenced to WRS-2, the Worldwide Reference System of paths and rows.

---

<sup>16</sup> <http://www.landsat.org>

On May 31, 2003, unusual artefacts began to appear within image data collected by the ETM+ instrument onboard Landsat 7. The problem was caused by failure of the Scan Line Corrector (SLC), which compensates for the forward motion of the satellite. In response to the SLC anomaly, the USGS has developed several new products to improve the utility of Landsat 7 data captured with the non-functioning SLC.

The standard processing chain of Landsat data includes a fully geometric, radiometric and atmospheric correction. The data is freely distributed, via a number of websites, e.g. the USGS earth explorer<sup>17</sup>.

Landsat data have been used by government, commercial, industrial, civilian, military, and educational communities throughout the world. The data support a wide range of applications in such areas as global change research, agriculture, forestry, geology, resource management, geography, mapping, water quality, and oceanography.

### 3.1.2 Sentinel

In the year 1998 the European Commission (EC) and the European Space Agency (ESA) initiated the Global Monitoring for Environment and Security (GMES) program. It is based on data received from Earth observation satellites and ground-based networks. Since 2007, as part of the GMES Space Component, ESA is developing five series of operational satellites, called Sentinels missions. ESA provides open and free access to the data collected by the Sentinel satellites. No distinction is made between public, commercial and scientific use, nor between European and non-European users. Thru the Copernicus Open Access Hub<sup>18</sup> data from all sentinel missions may be downloaded. Another and very easy source is the EO Browser<sup>19</sup>. In addition, in March 2016 ESA reached an agreement with NASA, NOAA and USGS on the use of data. Since then, these agencies have been allowed to transfer the data and integrate it into their existing database systems.

SENTINEL-1 is especially designed to help monitor sea ice, marine environments, land surface motion etc. with a set of two polar-orbiting RADAR satellites operating day and night. The SENTINEL-1 satellites were launched in April 2014 and April 2016 respectively. The C-Band Synthetic Aperture Radar (SAR) of the two satellites offers a 6-day repeat cycle at the equator. The radar can operate in four different observation modes (resolutions range  $\times$  azimuth):

- Strip-Map-Mode: 80 km wide stripes with a resolution of 5 $\times$ 5 m
- Interferometric wide swath mode: 250 km wide stripes with a resolution of 5 $\times$ 20 m
- Extra-wide swath mode: 400 km wide stripes with a resolution of 20 $\times$ 40 m
- Wave mode: individual areas of 20 $\times$ 20 km with a resolution of 5 $\times$ 5 m

SENTINEL-2 is a multi-spectral imaging mission, consisting of two polar-orbiting satellites (2A/2B) at the same sun-synchronous orbit, designed with a focus on the monitoring of vegetation, soil and water cover, as well as the observation of inland waterways and coastal areas. The Multispectral Instrument (MSI) samples 13 spectral bands: four bands with 10 m, six bands with 20 m and three bands with 60 m GSD. Compared to Landsat 7 / 8 / 9 the MSI of Sentinel 2 has some distinct differences, see Figure 61. In order to provide more detailed information about vegetation properties, three narrow red edge channels were added (Channel 5 – 7). Channel 1 is added to determine the aerosol concentration; Channel 9 is added to obtain information about the water vapor in the atmosphere. Together with channel 10, which is able to detect cirrus clouds all necessary data is collected for automatic atmospheric correction. Other than Landsat Sentinel 2 has no thermal channels.

---

<sup>17</sup> <https://earthexplorer.usgs.gov/>

<sup>18</sup> <https://scihub.copernicus.eu/>

<sup>19</sup> <https://apps.sentinel-hub.com/eo-browser/>

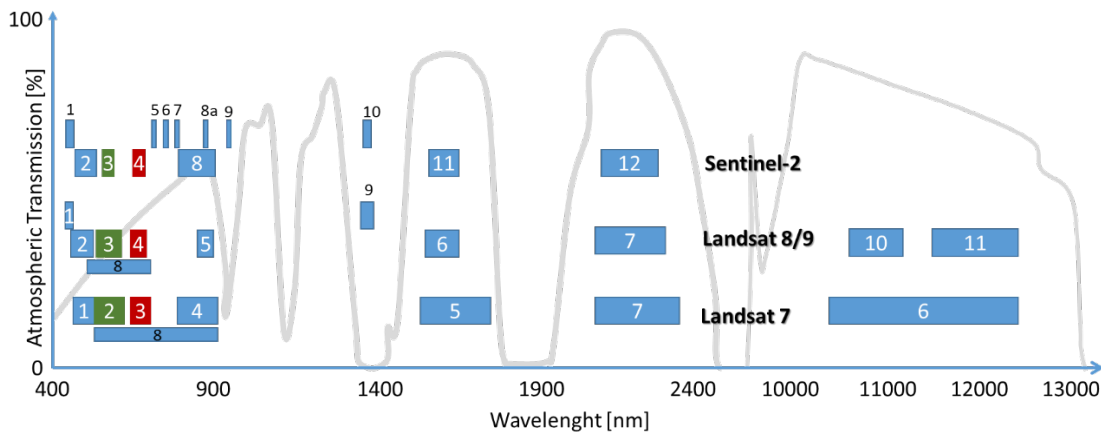


Figure 61: Comparison of spectral channels (bands) of Landsat 7 / 8 / 9 and Sentinel 2, Gray Background

The maximum continuous acquisition of an image from one SENTINEL-2 satellite is 15.000 km. The continuous acquisition is called a "data-take", and the data-take forms the base of the subsequent product tree. If a data-take is acquired by two separate receiving stations, the data-take may be sub-divided into datastrips. When the satellite is switched from one observation mode to another, a datastrip may include several distinct observation segments separated by gaps of an integer number of granules.

After the reception of the data at the ground station, a fully automated workflow processes the raw data from the satellite. Figure 62 illustrates the workflow. At the end, the user may obtain so called Level 1C data, which refers to a geometrically corrected top-of-atmosphere reflectance image.

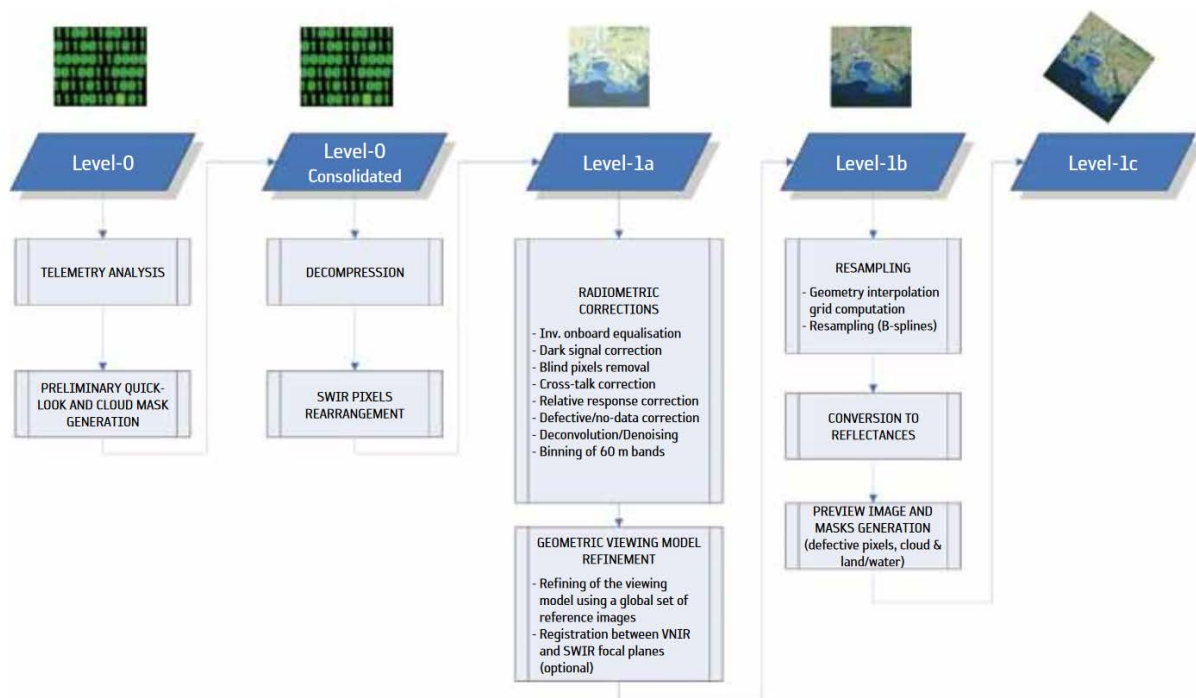


Figure 62: Postprocessing workflow of Sentinel 2 raw satellite data at ESA ground stations

For multitemporal data analysis and change detection an atmospherically corrected image is required. Since March 2018 the atmospheric correction and thus the Level 2A generation has also become a standard ESA product, extending the above shown postprocessing workflow. Formerly it could only be done by the user via the freely available Sentinel 2 toolbox SNAP<sup>20</sup>. SNAP, by the way is an advanced tool for the analysis of satellite remote sensing data. The main output of the Level-2A granules are Bottom-Of-Atmosphere (BOA) corrected reflectance ortho images. Additional outputs are an Aerosol Optical Thickness (AOT) map, a Water Vapour (WV) map and a Scene Classification Map (SCM) together with Quality Indicators (QI) for cloud and snow probabilities at 60 m

<sup>20</sup> <https://sentinel.esa.int/web/sentinel/toolboxes/sentinel-2>



resolution. Level-2A output image products are resampled and generated with an equal spatial resolution for all bands (10 m, 20 m or 60 m).

The underlying Sen2Cor processor algorithm for the atmospheric correction is a combination of state-of-the-art techniques adapted to the Sentinel-2 environment together with a scene classification module. The scene classification algorithm detects clouds, snow and cloud shadows and generates a classification map, which consists of three different classes for clouds (including cirrus), together with a basic land cover classification with six different classes (shadows, cloud shadows, vegetation, not vegetated, water and snow).

SENTINEL-3 is mainly designed for ocean observation and the measurement of land- and sea surface temperatures as well as the monitoring of the ice topography in the arctics. Sentinel-3A was launched on 16 February 2016 and Sentinel-3B on 25 April 2018. The payload of the Sentinel 3 satellites mainly consists of the following five instruments:

- OLCI (Ocean and Land Colour Instrument)
- SLSTR (Sea and Land Surface Temperature Radiometer)
- SRAL (Sentinel-3 Ku/C Radar Altimeter)
- MWR (Microwave Radiometer)
- POD (Precise Orbit Determination)

SENTINEL-4 is planned for 2021. The objective of the future Sentinel-4 mission is to monitor key air quality trace gases and aerosols over Europe at high spatial resolution with a fast (hourly) revisit time in support of the GMES Atmosphere Service<sup>21</sup>.

SENTINEL-5 is planned for 2021 /2022. Focus is an operational atmospheric monitoring mission of trace gas concentrations for atmospheric chemistry and climate applications. The Sentinel-5 mission is a payload, consisting of a single instrument named UVNS; it will be hosted as a CFI (Customer Furnished Item) on a post-EPS (MetOp) spacecraft, i.e. MetOp-SG-A1, and will be operated by EUMETSAT.<sup>22</sup>

SENTINEL-6 is planned for 2020 and 2026 and also known as Jason-CS satellite. It will carry a radar altimeter package to continue the high-precision, low-inclination altimetry missions of Jason-2 and -3. It will complement the high-inclination measurements on Sentinel-3 to obtain high-precision global sea-surface topography for the marine and climate user community<sup>23</sup>.

There is a wide range of Earth observation satellites other than Landsat and the Senstinels in orbit which will not be discussed here in detail, please refer to overviews in the internet, e.g. Wikipedia<sup>24</sup>.

### 3.1.3 High spatial and temporal resolution Earth observation satellites

Since the start of the **IKONOS** satellite in 1999 with a ground resolution of 1 m a new era of Earth observation has been started. These so called *1m-satellites* provide panchromatic data with a ground resolution of one meter or better.

**WorldView-3** is a multi-payload, multi-spectral, high-resolution commercial satellite sensor operating in a sun-synchronous orbit at an altitude of 617 km. The space craft is 5.7 m tall and weighs approximately 2.200 kg. At nadir WorldView-3 satellite provides 31 cm panchromatic resolution, 1.24 m multispectral resolution, 3.7 m short wave infrared resolution (SWIR). Additionally multispectral data to detect aerosols, water, desert clouds, snow etc. with a GSD of 30 m is collected, Figure 63. The swath width at nadir is 13.1 km.

<sup>21</sup> <https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/copernicus-sentinel-4>

<sup>22</sup> <https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/copernicus-sentinel-5>

<sup>23</sup> <https://directory.eoportal.org/web/eoportal/satellite-missions/j/jason-cs>

<sup>24</sup> [http://en.wikipedia.org/wiki/List\\_of\\_Earth\\_observation\\_satellites#Earth\\_Observing\\_System](http://en.wikipedia.org/wiki/List_of_Earth_observation_satellites#Earth_Observing_System)

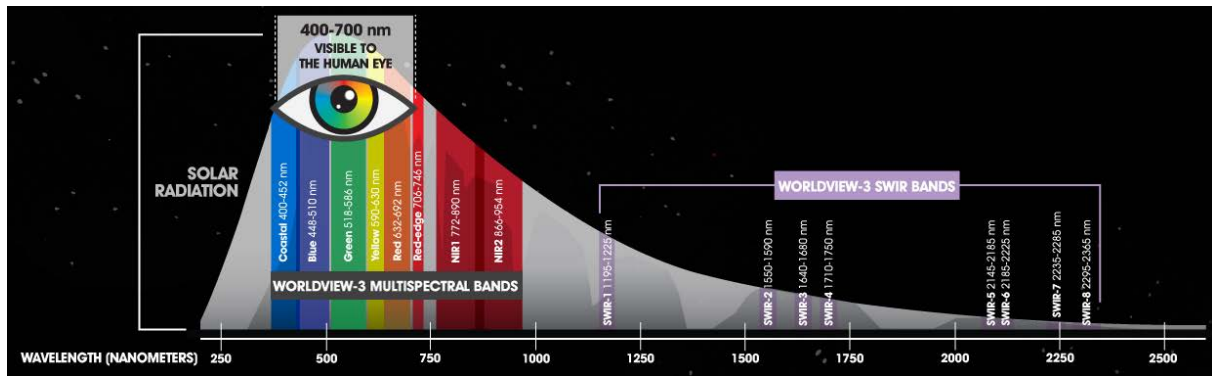


Figure 63: Spectral properties of Worldview 3 satellite, Source: www.digitalglobe.com

With its very precise and rapid pointing options, can collect data from a larger area. The satellite has an average revisit time of <1 day, applying 30° oblique view. For 20° or less off nadir images revisit rate goes down to 4.5 days. The satellite is capable of collecting up to 680,000 km<sup>2</sup> of data per day. The geo-location accuracy of the images is predicted to be less than 3.5 m CE90 without ground control. The successor Worldview 4 was lost, after 3 years in the Orbit in January 2019, because the control moment gyros on WorldView-4, have failed, preventing the spacecraft from pointing accurately.

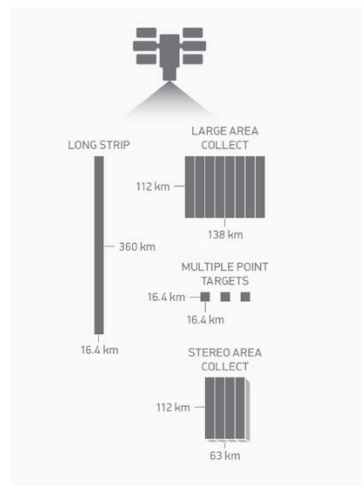


Figure 64: Tasking options of a high resolution EO-Satellite, Worldview 2

Modern commercial satellites are able to collect data according to the very concrete customer requests. It means that - unlike government satellite missions whose satellites follow consistent paths (Landsat, Sentinel 2) – commercial high resolution satellites can be tasked to capture a certain location at a certain time. Complex tasking of the satellite allows for the coverage of a larger area to compensate for the small footprint of a single image, see Figure 64.

It guarantees the availability of satellite data, which otherwise could be missing or delayed due to the change in the satellite's route. This ability has radically changed the disaster response and mitigation: Whenever a natural or man-made disaster occurs, high-resolution satellites are the first to provide a remote detailed view of the damaged territories inaccessible from the ground.

Beside the various advantages of high resolution satellite data, there are three drawbacks to consider:

- The aerial coverage of a single satellite scene is small (e.g. 10 x 10 km), compared to Landsat or Sentinel 2. Therefore, these high resolution satellites are most suitable for many local or regional applications, rather than national mapping projects.
- The development, launch and operation of commercial satellites is not cheap, therefore the specially tasked satellite data is not cheap. A single scene of 15 x 15 km may cost several thousand Euros.
- The lack of regular image coverage and a fairly small time frame available (since 2010) makes high-resolution satellite imagery data less suitable for multitemporal change detection analysis, compared to open source imagery.

The US-company **Planet labs** follows a different strategy, they want to provide satellite data with a very high temporal resolution, using a large number of small and relatively cheap satellites. The focus of the satellite system is agriculture and forestry as well as intelligence, security and infrastructure. The **PlanetScope** satellite

constellation consists of multiple launches of groups of individual satellites. Therefore, on-orbit capacity is constantly improving in capability or quantity, with technology improvements deployed at a rapid pace. Each PlanetScope satellite is a CubeSat 3U form factor (10 cm by 10 cm by 30 cm). The complete PlanetScope constellation of approximately 130 satellites is able to image the entire land surface of the Earth every day (equating to a daily collection capacity of 200 million km<sup>2</sup>/day). See Figure 65 for the simultaneous data acquisition of the satellite network.

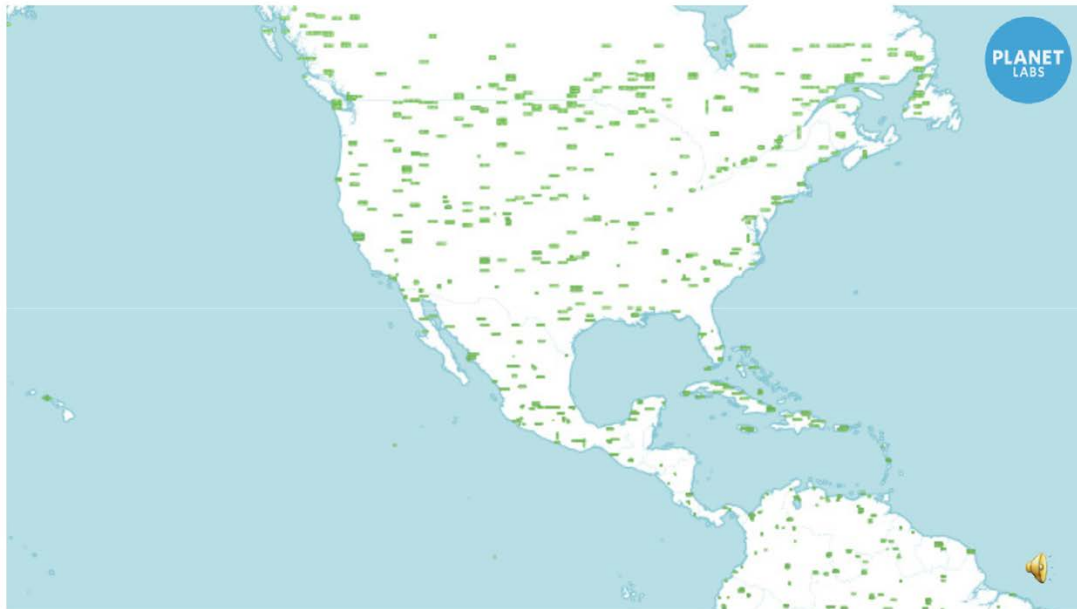


Figure 65: Simulations data acquisition of Planet labs  $\mu$ -satellites, Source: Planet Labs (<https://www.planet.com/>)

#### 3.1.4 Availability and Prices

Nowadays the standard pre-processing of satellite data is highly automated, enabling a GIS-ready accessibility via the internet within a few hours after acquisition. The price of remote sensing data is dependent upon three driving factors:

- *Urgency* - emergency services - the faster you need it, the more you pay
- *Age of the data* - the more recent the data, the higher its value
- *Spatial resolution* - the higher the spatial resolution, the higher the cost

The price structures of satellite data are generally complicated, including many clauses and pitfalls. Also several constraints apply for data distribution, because the user generally obtains limited rights to the data. Airborne remote sensing

The price structures of satellite data are generally complicated, including many clauses and pitfalls. Also several constraints apply for data distribution, because the user generally obtains limited rights to the data.

#### 3.1.5 Digital airborne frame camera

Even though satellite remote sensing provide data with a ground sampling distance (GSD) of 0.6 m, many applications call for much higher ground resolution to be used in mapping scales of 1: 200 – 1:10.000. Aerial photogrammetry involves the use of photographs taken in a systematic manner from the air. They are then controlled by land survey and measured by photogrammetric techniques. The accuracy of aerial photogrammetry is comparable with those obtained by land survey, and in many cases the work is carried out more economically. Modern high resolution airborne camera systems are multi head camera systems with a footprint of up to 450 Megapixels, acquiring images in RGB and Near-infrared (NIR). The cameras are equipped with a forward motion compensation to suppress image blur due to the fast flying aircraft. The cameras are fully calibrated and combined with an inertial measurement unit (IMU) to determine the precise exterior orientation of the cameras on the fly. Based on this information a fully automated geoprocessing (= direct georeferencing) without the need of further ground control points is possible.



Figure 66: Photogrammetric Multi Head Camera System - Voxel UltraCam with 3 panchromatic and 4 colour camera heads<sup>25</sup>

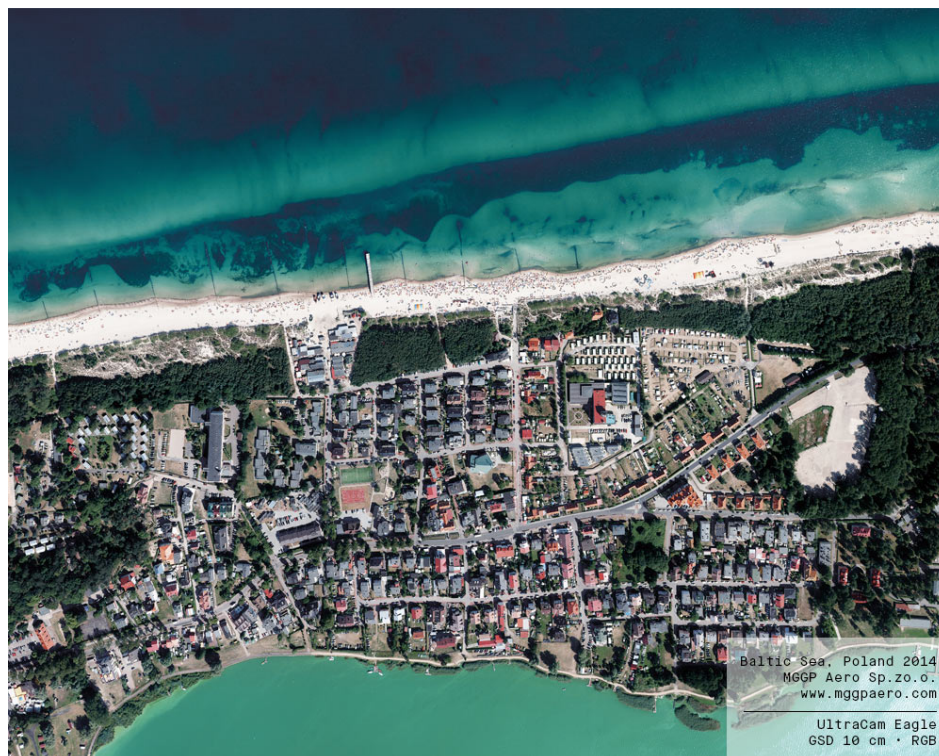


Figure 67: Image example of an airborne image, acquired with UltraCam Eagle camera<sup>26</sup>

### 3.1.6 UAS

UAS (Unmanned Airborne Systems), commonly known as drones are able to close the large gap between terrestrial and airborne and satellite-based geodata acquisition. The carrier platforms on the market are very diverse and can be equipped with a wide variety of sensors to capture image and geodata. The data processing process has been considerably simplified and accelerated in recent years. This makes it possible to generate very flexible image-based three-dimensional geodata of small areas and to measure them with high precision.

Due to the low costs of flying and the low weather dependency, processes and changes can be documented, evaluated and analyzed, such as the construction progress on a construction site, the growth of plants, the development of weather phenomena or even a volcanic eruption. With the high ground resolution and the high achievable 3D accuracies of a few centimeters, UAS can also be used for a variety of surveying tasks. In this context, the combination of terrestrial and airborne measurement techniques, i.e. terrestrial laser scanning and UAS point clouds, offers additional value. The possibility to fly under the cloud cover or to use the chance of a small cloud gap increases the possible flight hours many times compared to classical airborne aerial photography. This increases reliability for customers and the opportunity to generate new services.

<sup>25</sup> <https://www.vexcel-imaging.com/>

<sup>26</sup> <https://www.vexcel-imaging.com/media/>

In addition to common digital still and video cameras, a wide range of experience has also been gained with other imaging sensors, such as laser scanners, multi-spectral cameras, thermal infrared cameras and hyperspectral sensors. In addition, attempts are being made to use UAS as multi-sensor platforms, for instance to combine a digital camera, two laser scanners, a spectrometer and an infrared camera for forest inventory. A UAS can also be understood as a universal sensor platform and can integrate further specialized sensor technology, e.g. from the fields of geophysics and meteorology, such as sensors for temperature, pressure, humidity, wind, CO<sub>2</sub>, flow and magnetic field measurement. However, these non-imaging sensors will not be discussed further, as they are not used in remote sensing.

Despite the highly automated workflow, the user has a great influence on the achievable accuracies, e.g. by choosing the appropriate camera, intelligent flight planning and ground control point configuration.

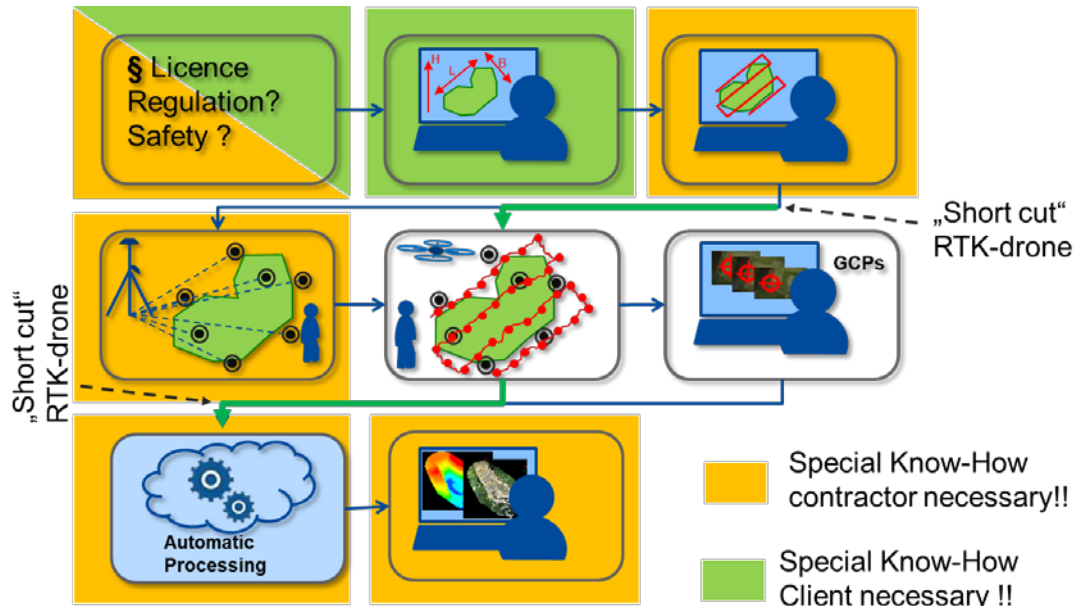


Figure 68: Drone based geo data acquisition workflow

There is a high number of drones available on the market in the < 5 kg weight class. The offer ranges from inexpensive systems, which are also used by hobbyists, to professional drones optimized for special applications. Nevertheless, different categories and classes can be distinguished. Firstly, there is the flight principle (*multicopter* or *fixed wing*), secondly the payload and the sensor technology used and thirdly functions and software for safety, flight planning and execution.

While in aerial photogrammetry the first question is which camera model is used, in a UAS the first question to be answered is the suitable carrier platform. Various aircraft are available as UAS carrier platforms for aerial photography: fixed wing model aircraft, model helicopters, quadcopters or balloons / blimps, see.



Figure 69: Different Drone types

At first glance, fixed wing or model aircraft seem to be ideally suited as a carrier platform for spatial data acquisition. After all, model airplanes can fly 100 km/h or faster without any problems and stay in the air for one or more hours, depending on the engine. These advantages predestine aeroplanes for applications outside the visibility of the pilot, which is still prohibited in many parts of the world. Due to the high airspeed, wing flyers must always have a certain distance to the object or maintain a minimum flight altitude. Otherwise, motion blur can hardly be avoided. With low flight altitudes and longitudinal overlaps of 80% and more, the image sequence time can also become a limiting factor in order to save all images. Depending on the take-off weight of the drone, a manual take-off (max. weight 5 - 6 kg) is feasible or a take-off and landing runway is necessary, which must be taken into account during flight preparation. Taking all of the above mentioned into account, fixed wing drones are ideally suited for agricultural or forestry applications, but not for any kind of urban mapping.

Multicopters have four or more rotors. Quadcopters are the most common type of aircraft with four propellers. The big advantages of multi-copters are their ability to stop at one spot or to fly at a defined speed. The modern self-controlling electronics make multicopters very easy to operate and fly. Multicopters are usually driven by electric motors, so they are quiet and can remain in the air for between 10 and 30 minutes, depending on their size. From a photogrammetric point of view, the advantages lie above all in the possibility of controlled systematic flying and the possibility of combining vertical and oblique images.

The advantages of model helicopters are comparable to those of multicopters in terms of flight aspects. In addition, there is the good ratio between flight weight and payload. Larger model helicopters are therefore ideally suited, for example, to get a laser scanner into the air. However, flying a model helicopter requires a greater degree of experience. They are usually powered with gasoline and are therefore quite loud and always require an individual permit. This means that they are not suitable for regular, flexible flight operations, but only for special applications in which a large payload is required, e.g. for geomagnetics measurements or the measurement of wind turbulence in one or more wind turbines.

Blimps are airships without internal scaffolding. Equipped with several rotors, they can stand long, sometimes several days, above a location or observe and track objects at low speed. Today, helium is used as the carrier gas, which must either be discharged or recovered during transport. The biggest disadvantage of a blimp is its susceptibility to wind due to its large surface area and the costly transport that requires a small trailer or transporter. Due to their weight (usually more than 5 kg), an individual permit is required for commercial operation.

### 3.1.7 Flight planning

For systematic aerial surveys the autopilot has to take over the flight control. This ranges from manually assisted take-off of a given flight path along the waypoints to automatic take-off and landing, also in case of loss of contact between ground station and autopilot, to (future) completely automatic flight path generation on the fly including the detection and avoidance of obstacles.

For flight mission planning, the flight trajectory shall be determined in advance. This is done on the basis of maps, orthophotos or other geoinformation in 2D or 3D by defining so-called waypoints in a global coordinate reference system (e.g. WGS'84). The final flight planning is usually done on site in order to determine the concrete place of starting and landing and also to ensure that no obstacles lies in the planned trajectory and that safety distances can be maintained. Nevertheless, a preliminary planning in the office is very useful in order to estimate the flight effort,

the number of photos and the necessary number of ground control points. Within the flight planning the ground resolution and the geometric accuracy of a flight are defined.

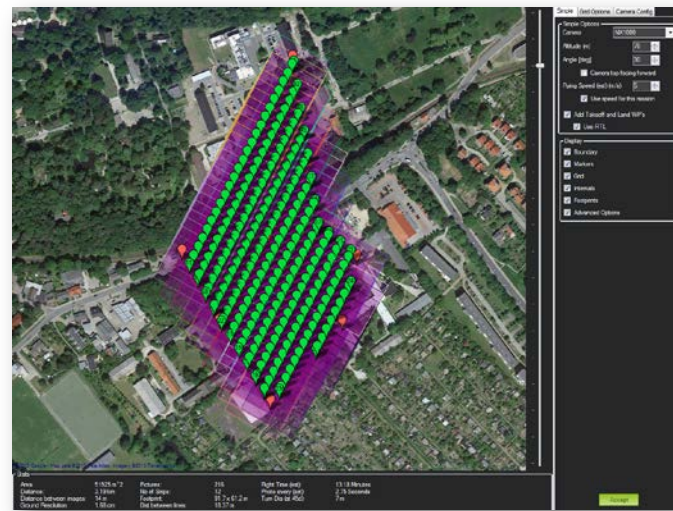


Figure 70: Planned flight path for drone survey

During the drone flight, the earth's surface is captured redundantly in order to obtain 3D coordinates from the two dimensional images. For drone flights, longitudinal overlaps of 80 % and transverse overlaps of 60 % are common. For heavily fissured surfaces, a longitudinal overlap of 90 % and a transverse overlap of 80 % can be useful. An increase in the transverse overlap always means increased flight effort and, of course, higher costs for data processing. These high overlaps are necessary for the so-called "Multi Ray Matching" and guarantee a consistently high quality of the results with minimum number of ground control points (GCP). GCP's are used for precise geometric positioning. GCP'S are usually measured with geodetic GPS-receivers or total stations before the flight and signaled on the ground in the form of plates or markings. The GCP's should preferably at the border of the survey area. The minimum number of GCP's is 5, the optimal number depends on the size and the form of the block and other factors. The size or diameter of the control points depends on the ground resolution. In general, it should be 5 - 10 times the soil resolution and easily recognizable in the landscape, see Figure 70.



Figure 71: Ground control points for every situation

### 3.1.8 UAS Photogrammetry

The simultaneous (external) orientation of several images is a central task of photogrammetry. Therefore, neighboring images have to "know" each other thru common thru image matching. This is done with SIFT- and SURF-algorithms, which identify a large number features in the images and match them in adjacent images. For image orientation in a photogrammetric sense, the Structure from Motion (SfM) method is combined with bundle adjustment including a self calibration procedure. After the orientation of the images the precise position and the viewing direction (= exterior orientation) of each camera is known. Thereafter a dense point cloud of the surface is generated which is used for further ortho photo generation. As a result of the photogrammetric processing, which

requires little manual input, three basic products are generated. A digital orthophoto, a digital surface model (DSM) and a 3D-point cloud.

### Digital Orthophotos

A digital orthophoto has the same properties as a map and can be used to measure true distances. It is an accurate representation of the Earth's surface. To create a digital orthophoto, several key input files are necessary: aerial digital images, orientation information of the images in space, and a digital surface model (DSM, or a digital terrain model (DTM)), a 2.5-dimensional description of the Earth's surface. The results of the bundle adjustment include a camera calibration (interior orientation) and the perspective centre of the camera (exterior orientation). At a minimum, the DEM can be a regularly spaced grid of measured points, each containing an x, y and z value. A more robust digital terrain model (DTM) can also be used which includes strategically placed measurement points, dense breaklines, and ridgelines. In the rectification process, the errors due to the terrain displacement are corrected and single images are combined into an orthophoto mosaic.

The limitations of ordinary digital orthophotos are that expansion features, such as bridges, create problems. DTM data is captured at ground level, so bridges that are rectified with this data are "pulled down to the ground," giving them a distorted appearance. Elevated features (e.g., buildings, trees, power lines) also create a problem due to radial displacement. Distortion increases with the distance from the centre of the aerial photograph-features, such as buildings, lean noticeably at the corner of the images. The amount a feature leans in the final orthophoto depends on the percentage of overlap in the aerial photography and the height of that feature. The higher the percentage of overlap in the aerial photography used, the less features will lean because the amount of photography used from the outer edge in the Orthophoto mosaic is reduced. This distortion can have an impact on the functional and aesthetic features of a digital orthophoto.

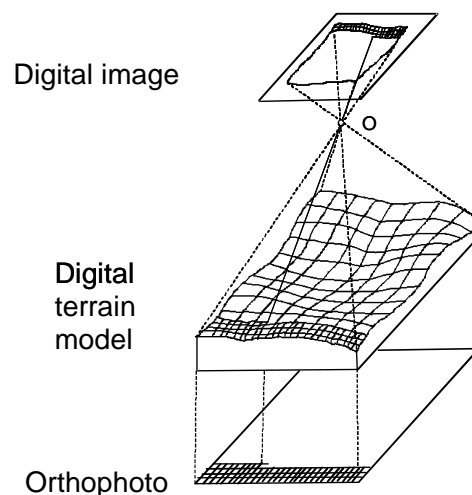


Figure 72: Digital orthophoto process

True orthoimagery, which are common for drone imagery may overcome these problems. Therefore, the DSM has to represent the buildings, bridges etc. The end- and side-overlap of the aerial images necessary has to be more than 50 %, because when a building is repositioned correctly, blind spots occur. These hidden spots have to be filled with information of other images, thus requiring an intelligent mosaicing procedure, see Figure 73.



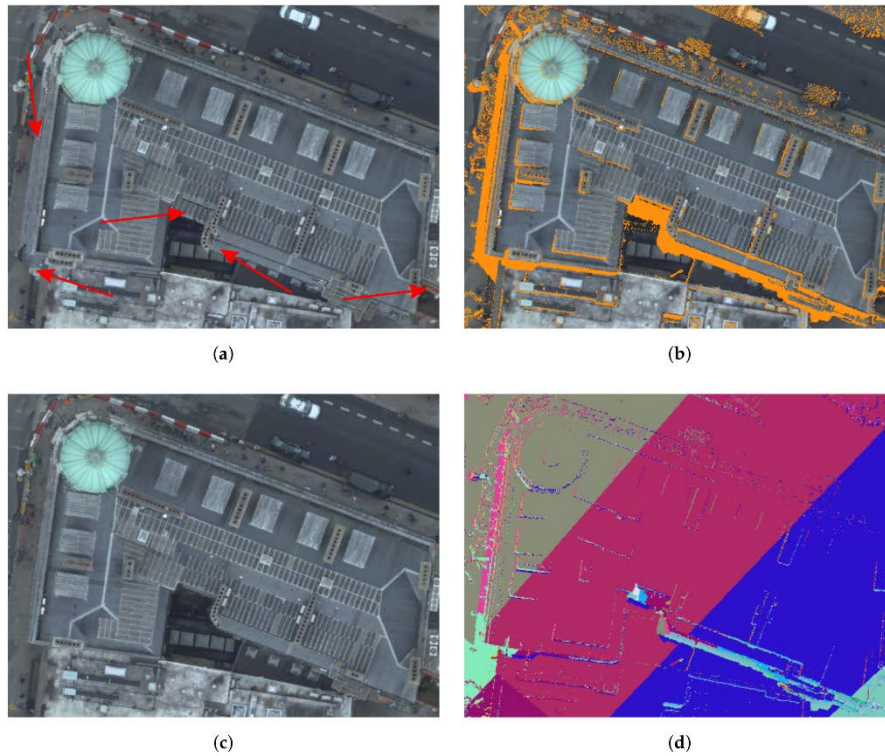


Figure 73: (a) Differentially-rectified orthophoto, (b) True orthophoto with occluded areas shown in orange, (c) True orthophoto generated using the closest visible image criteria, and (d) Contribution of images to the true orthophoto shown in part (c). Pixel size is 10 cm. Source: Gharibi and Habib, 2018

## 4 Remote sensing applications

Since there are so many different remote sensing applications in various disciplines, a complete coverage cannot be given. Since (satellite) remote sensing enables the measurement of certain parameters, Table 6 may give a short overview.

As the satellites are constantly acquiring images of the earth surface and the data processing chains are highly automated, more and more **services** are generated to monitor all kinds of changes and developments. For instance on the basis of Sentinel satellites within the EU Copernicus program, services about marine, land, climate change, security etc. are operational<sup>27</sup>. The Copernicus Land Monitoring Service (CLMS) for instance provides GIS data on land cover and its changes, land use, vegetation state, water cycle and earth surface energy variables for users in Europe and across the World in the field of environmental terrestrial applications.

<sup>27</sup> <https://www.copernicus.eu/en>

Discipline	Measurement	Discipline	Measurement
Atmosphere	Cloud Properties	Ocean	Surface Temperature
	Radiative Energy Fluxes		Phytoplankton
	Precipitation		Dissolved Organic Matter
	Tropospheric Chemistry		Surface Wind Fields
	Stratospheric Chemistry		Ocean Surface Topography
	Aerosol Properties	Cryosphere	Land Ice Change
	Atmospheric Temperature		Sea Ice
	Atmospheric Humidity		Snow Cover
Lightning			
Land	Land Cover/Land Use Change	Solar Radiation	Total Solar Radiation
	Vegetation Dynamics		Ultraviolet Spectral Irradiance
	Surface Temperature		
	Fire Occurrence		
	Volcanic Effects		
	Surface Wetness		

Table 6: Overview of remotely sensible properties by satellite remote sensing

## 5 Basics of digital image processing / image analysis

Using radio waves, data from Earth-orbiting satellites are transmitted on a regular basis to properly equipped ground stations. As the data are received they are translated into a digital image (raster data) that can be displayed on a computer screen. A **pixel** (Picture Element) is the smallest element of an image (IFOV, Instant...). The *geometric resolution* of an image is generally determined by the size of a pixel on the ground, e.g. 5 m on a SPOT 5 scene. Similarly the expression ground sampling distance (GSD) is used. A **mixel** is a pixel in which different object signatures, e.g. house and surroundings are represented. Mixels cause different problems in the classification procedure. The proportion of mixels depends upon the geometric resolution and the spatial distribution of the objects.

### 5.1 Filtering

For many remote sensing Earth science applications, the most valuable information that may be derived from an image is contained in the *edges* surrounding various objects of interest. *Edge enhancement* delineates these edges and makes the shapes and details comprising the image more conspicuous and perhaps easier to analyse. Edges may be enhanced using either *linear* or *nonlinear edge enhancement* techniques.

### 5.2 Pre-processing operations

*Remote sensing systems do not function perfectly.* Also, the Earth's atmosphere, land, and water are complex and do not lend themselves well to being recorded by remote sensing devices that have constraints such as spatial, spectral, temporal, and radiometric resolution. Consequently, error creeps into the data acquisition process and can degrade the quality of the remote sensing data collected. The two most common types of error encountered in remotely sensed data are *radiometric* and *geometric*.

- *Radiometric correction* attempts to improve the accuracy of spectral reflectance, emittance, or back-scattered measurements obtained using a remote sensing system (see section 2.1.1)
- *Geometric correction* is concerned with placing the reflected, emitted, or back-scattered measurements or derivative products in their proper planimetric (map) location so they can be associated with other spatial information in a geographic information system (GIS).

### 5.3 Multispectral classification

Multispectral classification may be performed using a variety of methods, including:

- algorithms based on *parametric* and *non-parametric* statistics that use ratio- and interval-scaled data and *non-metric* methods that can also incorporate nominal scale data,
- the use of *supervised* or *unsupervised* classification logic,
- the use of *hard* or *soft (fuzzy) set classification* logic to create hard or fuzzy thematic output products,
- the use of *per-pixel* or *object-oriented classification* logic,
- the use of *machine or deep learning algorithms*,
- *hybrid* approaches.

*Parametric* methods such as maximum likelihood classification and unsupervised clustering assume normally distributed remote sensing data and knowledge about the forms of the underlying class density functions.

*Non-parametric* methods such as nearest-neighbour classifiers, fuzzy classifiers, and neural networks may be applied to remote sensing data that are not normally distributed and without the assumption that the forms of the underlying densities are known.

*Non-metric* methods such as rule-based decision tree classifiers can operate on both real-valued data (e.g., reflectance values from 0 to 100%) and nominal scaled data (e.g., class 1 = forest; class 2 = agriculture).

In an *unsupervised classification*, the identities of land-cover types to be specified as classes within a scene are not generally known *a priori* because ground reference information is lacking or surface features within the scene are not well defined. The computer is instructed to group pixels with similar spectral characteristics into unique clusters according to some statistically determined criteria. The analyst then re-labels and combines the spectral clusters into information classes.

In a *supervised classification*, the identity and location of some of the land-cover types (e.g., urban, agriculture, or wetland) are known *a priori* through a combination of fieldwork, interpretation of aerial photography, map analysis, and personal experience. The analyst attempts to locate specific sites in the remotely sensed data that represent homogeneous examples of these known land-cover types. These areas are commonly referred to as *training sites* (or ground truth) because the spectral characteristics of these known areas are used to train the classification algorithm for a possible land-cover mapping of the remainder of the image. Multivariate statistical parameters (means, standard deviations, covariance matrices, correlation matrices, etc.) are calculated for each training site. Every pixel both within and outside the training sites is then evaluated and assigned to the class of which it has the highest likelihood of being a member.

Supervised and unsupervised classification algorithms typically use *hard classification* logic to produce a classification map that consists of hard, discrete categories (e.g., forest, agriculture). Conversely, it is also possible to use *fuzzy set classification* logic, which takes into account the heterogeneous and imprecise nature of the real world. In the past, most digital image classification was based on processing the entire scene pixel by pixel. This is commonly referred to as *per-pixel classification*.

*Object-oriented classification* techniques allow the analyst to decompose the scene into many relatively homogenous image objects (referred to as patches or segments) using a multi-resolution image segmentation process. The various statistical characteristics of these homogeneous image objects in the scene are then subjected to traditional statistical or fuzzy logic classification. Object-oriented classification based on image segmentation is often used for the analysis of high-spatial-resolution imagery (e.g.,  $.7 \times 0.7\text{m}$  SPOT Pléiades-HR and  $0.41 \times 0.41\text{m}$  GeoEye 1).

*Convolutional Neuronal Networks (CNN)*, commonly known as deep learning techniques follow a completely different strategy. CNN detects and extracts characteristics of input images using filters. The recognition of structures within the image is location-independent. First, the CNN recognizes simple structures such as lines, spots of color, or edges in the first layers. In the other layers, the Convolutional Neural Network learns combinations of these structures such as simple shapes or curves. More complex structures can be identified with each layer. The data is scanned and filtered again and again in the layers. In the last step, the results are assigned to the classes or objects to be recognized.

Technically, deep learning CNN models are developed in four steps. First a definition phase, in which the objects of interest are defined, e.g. to detect palm trees or birds. Second, a training phase in which training data is developed. Thereby small image snippets with the relevant objects are prepared. CNN generally requires a lot of training data, which allows the network to separate the objects of interest from all other objects in the scene. The third step is each input image will pass it through a series of convolution layers with filters (Kernels), Pooling, fully connected layers (FC) etc. to classify an object with probabilistic values between 0 and 1.

In more detail, Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. Convolution of an image with different filters can perform operations such as edge detection, blur and sharpen by applying filters. Pooling is an example-based discretization process. The aim is to scan an input representation (image, hidden layer output matrix, etc.), reduce the dimensionality and make assumptions about the features contained in the subregions. Pooling reduces the number of parameters to be learned and thus the computational effort. The classifier is the last

step in a CNN. This is called a dense layer, which is a common classifier for neural networks. The dense layer is a downward layer from the pooling layer. In this layer, each node is connected to each node in the previous layer. Like any classifier, it needs individual features. So it needs a Feature Vector. Therefore, the multidimensional output from the convolutions has to be converted into a one-dimensional vector. This process is called "flattening". The last and final step is an evaluation of the results in order to access the classification accuracy and the quality of the training data as well as the (self learning) classifier. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.

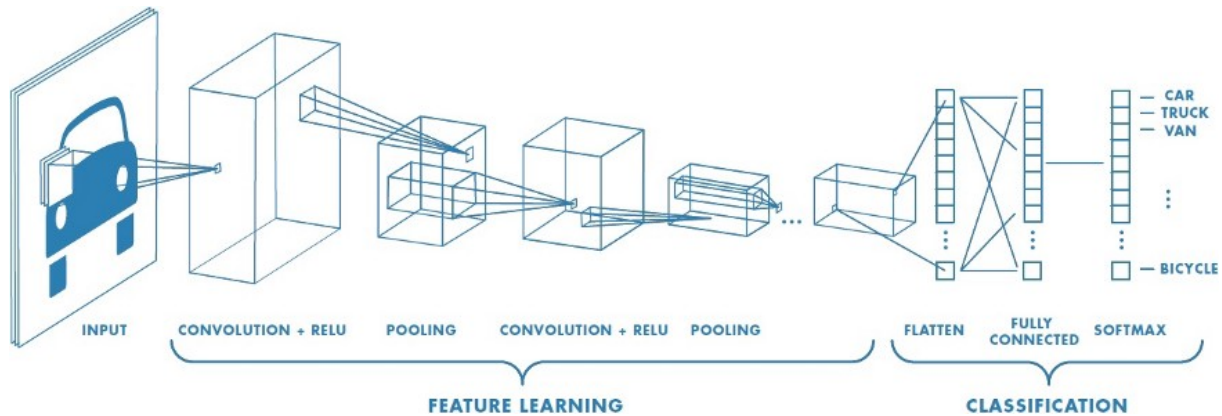


Figure 74: Complete flow of CNN for learning processing and classification<sup>28</sup>

*No pattern classification method is inherently superior to any other.* The nature of the classification problem, the biophysical characteristics of the study area, the distribution of the remotely sensed data (e.g., normally distributed), and *a priori* knowledge determine which classification algorithm will yield useful results. However we should have a healthy scepticism regarding studies that purport to demonstrate the overall superiority of a particular learning or recognition algorithm.

### 5.3.1 Multi spectral land use classification

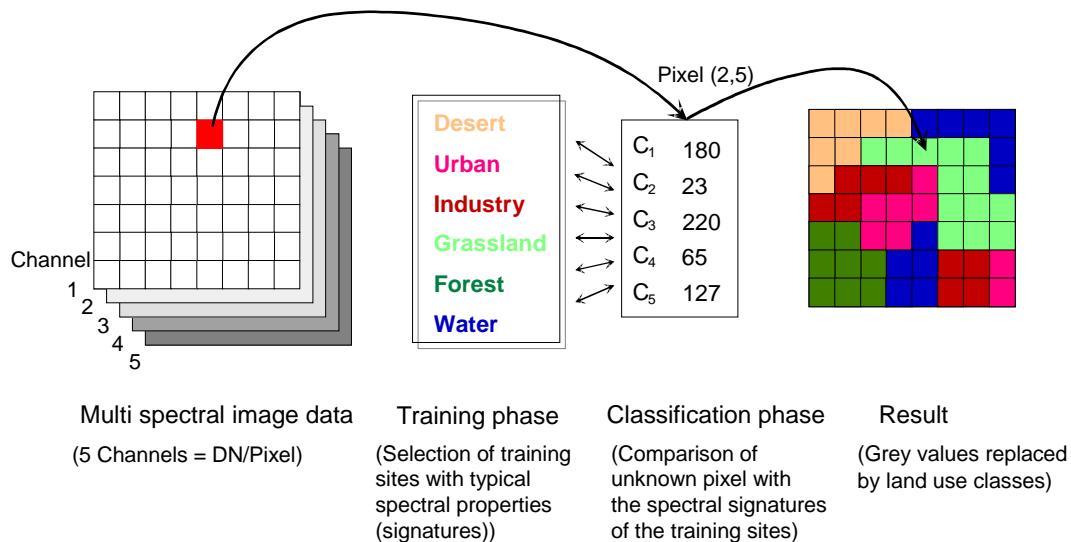


Figure 75: Procedure of supervised multispectral classification

The classification of satellite images with respect to the Earth cover was one of the first remote sensing applications. Thereby it is necessary to separate between:

- *Land cover* refers to the type of material present on the landscape (e.g., water, sand, crops, forest, wetland, man-made materials such as asphalt).
- *Land use* refers to what people do on the land surface (e.g., agriculture, commerce, settlement).

Prior to a classification a land-use and land-cover classification scheme has to be developed. All classes of interest must be selected and defined carefully to classify remotely sensed data successfully into land-use and/or land-

<sup>28</sup> <https://de.mathworks.com/solutions/deep-learning/convolutional-neural-network.html>

cover information. This requires the use of a *classification scheme* containing *taxonomically* correct definitions of classes of information that are organized according to logical criteria. If a hard classification is to be performed, then the classes in the classification system should normally be:

- *mutually exclusive*,
- *exhaustive*, and
- *hierarchical*.

*Mutually exclusive* means that there is no taxonomic overlap (or fuzziness) of any classes (i.e., deciduous forest and evergreen forest are distinct classes). *Exhaustive* means that all land-cover classes present in the landscape are accounted for and none have been omitted. *Hierarchical* means that sublevel classes (e.g., single-family residential, multiple-family residential) may be hierarchically combined into a higher-level category (e.g., residential) that makes sense. This allows simplified thematic maps to be produced when required.

It is also important for the analyst to realize that there is a fundamental difference between *information* classes and *spectral* classes.

- *Information classes* are those that human beings define.
- *Spectral classes* are those that are inherent in the remotely sensed data and must be identified and then labelled by the analyst.

Common land use classification schemas have been developed for different parts of the worlds, e.g. CORINE Land cover within the EU.

During the classification procedure an analyst may select *training sites* within the image that are representative of the land-cover or land-use classes of interest *after* the classification scheme is adopted. The training data should be of value if the environment from which they were obtained is relatively homogeneous. Each site is usually composed of many pixels. The general rule is that if training data are being extracted from  $n$  bands then  $>10n$  pixels of training data are collected for each class. This is sufficient to compute the variance–covariance matrices required by some classification algorithms.

There are a number of ways to collect the *training site* data, including:

- collection of *in situ* information such as forest type, height, percent canopy closure, and diameter-at-breast-height (DBH) measurements,
- on-screen selection of polygonal training data, and/or
- on-screen seeding of training data.

## 5.4 Digital change detection

Biophysical materials and human-made features on the surface of the Earth are inventoried using remote sensing and field techniques. Some of the data are fairly *static*; they do not change over time. Conversely, some biophysical materials and man-made features are *dynamic*, changing rapidly. It is important that such changes be inventoried accurately so that the physical and human processes at work can be more fully understood. In fact, it is believed that land-use/land-cover change is a major component of global change with an impact perhaps greater than that of climate change. *It is therefore not surprising that significant effort has gone into the development of change detection methods using remotely sensed data.*

Sometimes the *time period* selected over which change is to be monitored is too short or too long to capture the information of interest. *The analyst must be careful to identify the optimal change detection time period(s).* This selection is dictated by the nature of the problem. Traffic transportation studies might require a change detection period of just a few seconds or minutes. Images obtained monthly or seasonally might be sufficient to monitor the greening of a continent. Careful selection of the change detection time period can ensure that resource analysis funds are not wasted.

Most change detection studies have been based on the comparison of multiple-date *hard* land-cover classifications of remotely sensed data. The result is the creation of a *hard* change detection map consisting of information about the change in discrete categories (e.g., change in forest, agriculture). This is still very important and practical in many instances, but *we now recognize that it is ideal to capture both discrete and fuzzy changes in the landscape.*

Most digital image change detection is based on processing Date  $n$  and Date  $n + 1$  classification maps pixel by pixel. This is commonly referred to as *per pixel* change detection. *Object-oriented change detection* involves the comparison of two or more scenes consisting of many relatively homogenous image objects (patches or segments). The smaller number of relatively homogeneous image objects in the two scenes are then subjected to various change detection techniques.

Successful remote sensing change detection requires careful attention to:

- remote sensor system considerations, and
- environmental characteristics.

Failure to understand the impact of the various parameters on the change detection process can lead to inaccurate results. Ideally, the remotely sensed data used to perform change detection is acquired by a remote sensor system that holds the following resolutions constant: temporal, spatial (and look angle), spectral, and radiometric.

Two temporal resolutions should be *held constant* during change detection, if possible. First, use a sensor system that acquires data at approximately the *same time of day*. For example, Landsat TM data are acquired before 9:45 a.m. for Vietnam. This eliminates diurnal sun-angle effects that can cause anomalous differences in the reflectance properties of the remote sensing data. Second, acquire remote sensing data on *anniversary dates*, e.g., Feb 1, 2005, and Feb 1, 2007. Anniversary date imagery minimizes the influence of seasonal sun-angle and plant phenological differences that can negatively impact a change detection project.

Accurate spatial registration of at least two images is essential for digital change detection. Ideally, the remotely sensed data are acquired by a sensor system that collects data with the same *instantaneous field of view* on each date. For example, Landsat TM data collected at 30×30 m spatial resolution on two dates are relatively easy to register to one another.

It is possible to perform change detection using data collected from two different sensor systems with different IFOVs, for example, Landsat TM data (30×30 m) for Date 1 and Sentinel 2 data (20'20 m) for Date 2. *In such cases, it is usually necessary to decide on a representative minimum mapping unit (e.g., 20×20 m) and then resample both datasets to this uniform pixel size.* This does not present a significant problem as long as the image analyst remembers that the information content of the resampled data can never be greater than the IFOV of the original sensor system. Remotely sensed data used for change detection should be geometrically rectified to be within  $\pm 0.5$  pixel of its correct planimetric position.

Remote sensing systems like SPOT and QuickBird can collect data at off-nadir *look angles* as much as  $\pm 20^\circ$ ; i.e. from an *oblique* vantage point. Two images with significantly different look angles can cause problems when used for change detection. For example, a SPOT image of a maple forest acquired at  $0^\circ$  off-nadir will look directly down on the top of the canopy. Conversely, a SPOT image acquired at  $20^\circ$  off-nadir will record reflectance from the side of the canopy. Differences in reflectance from the two datasets may cause spurious change detection results. Therefore, the data used in remote sensing digital change detection should be acquired with approximately the same look angle, if possible.

Ideally, the same sensor system is used to acquire imagery on multiple dates. When this is not possible, the analyst should *select bands that approximate one another*. For example, Landsat MSS bands 4 (green), 5 (red), and 7 (near-infrared) and SPOT bands 1 (green), 2 (red), and 3 (near-infrared), can be used successfully with Landsat ETM+ bands 2 (green), 3 (red), and 4 (near-infrared). Many change detection algorithms do not function well when the bands in one image do not match those of the other image (e.g., utilising the Landsat TM band 1 (blue) with either SPOT or Landsat MSS data may not be wise).

Radiometric correction of multiple dates of imagery should be applied to perform change detection. Thereby a absolute radiometric correction which makes use of a model atmosphere in conjunction with *in situ* atmospheric measurements (if possible) to correct for path radiance is best.

- Most commonly relative radiometric correction is applied, either
  - Single image normalization using histogram adjustment, or
  - Multiple date image normalization using regression techniques.

The selection of an appropriate change detection algorithm is very important. First, it will have a direct impact on the type of image classification to be performed (if any). Second, it will dictate whether important “from-to” change information can be extracted from the imagery. Many change detection projects require that “from- to” information be readily available in the form of maps and tabular summaries.

*Change detection algorithms* commonly used include:

- write function memory insertion
- multi-date composite image
- image algebra (e.g., band differencing, band ratioing)
- post-classification comparison
- binary mask applied to date 2
- ancillary data source used as date 1
- spectral change vector analysis
- chi-square transformation
- cross-correlation
- visual on-screen digitization
- knowledge-based vision systems.

## 6 Recent developments and research issues

Recent advances in remote sensing and geographic information have paved the way for the development of hyperspectral sensors. **Hyperspectral remote sensing**, also known as imaging spectroscopy, is a relatively new technology that is currently being investigated by researchers and scientists with regard to the detection and identification of minerals, terrestrial vegetation, and man-made materials and backgrounds. Hyperspectral remote sensing combines imaging and spectroscopy in a single system which often produces large data sets that require new processing methods. Hyperspectral data sets are generally composed of about 100 to 200 spectral bands of relatively narrow bandwidths (5-10nm), whereas multispectral data sets are usually composed of only about 5 to 10 bands of relatively wide bandwidths (70-400 nm).

Based on satellite data, simple vegetation indices, multispectral classifications and knowledge-based classifications have been successfully implemented in a number of studies. For high resolution aerial images, such pixel-based methods show their limitations because they rely solely on the spectral properties of each pixel. To overcome this problem, **object-based classification** methods have been developed in the last couple of years. Such methods use a multi-level segmentation procedure to generate larger image elements before the classification. The classification procedure may use additional data sources and rules. Through membership functions it is possible to define fuzzy logic rules, after which certain classes may be formulated. The ability to incorporate knowledge about the neighbourhood of each segment, not only in the classification procedure but also in the segmentation phase, is another key difference to the conventional pixel-based classification approach.

**Oblique images** provide a new data source for photogrammetry and GIS. In the past oblique images were generally taken for visualisation and interpretation purposes, rather than for metric applications. An exception is the military sector where oblique images have long been used for reconnaissance purposes. Oblique images were therefore generally outside of the focus of photogrammetrists.

The use of standard vertical orthoimages as a topographic background in a GIS is nowadays very common, thus generating a strong demand for current photogrammetric airborne and high resolution satellite data. Planners, administrative users and the general public use the available orthoimages, e.g. in Google Earth and other similar services mainly for orientation and visual inspection of selected features. Yet vertical orthoimages may not be easily interpreted by everyone. Due to the intuitive use of oblique images, which are similar to the common human perspective, these images are very attractive to decision makers, as well as for the general public. To fully exploit the information from the oblique perspective, a minimum of four images from all sides have to be acquired and managed. Standard GIS packages do not support oblique images, due to their geometry with varying scales, therefore new viewers and software packages need be developed to guide the users and provide them with the necessary functionality.

## References

- Campbell, J.B. (2007): Introduction to Remote Sensing.- 4Rev Ed., 546 p.; New York, London
- Canty, M.J. (2019): Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for Python, Fourth Edition, 530 p.
- Gharibi, H. and Habib, A. (2018): True Orthophoto Generation from Aerial Frame Images and LiDAR Data: An Update.- Remote Sens. 2018, 10(4), 581; <https://doi.org/10.3390/rs10040581>
- Huete, A. R. (1988): A soil-adjusted vegetation index (SAVI), Remote Sensing of Environment, 25, 53-70.
- Huete, A., Justice, C., Liu, H. (1994): Development of vegetation and soil indices for MODIS-EOS.- Remote Sensing of Environment. Vol. 49, no. 3, pp. 224-234.
- Lillesand, T. M., Kiefer, R. W., Chipman, J. (2015): Remote Sensing and Image Interpretation.- 7. ed.: 768 p.; John Wiley & Sons, NewYork
- Jensen, J.R. (2013): Remote Sensing of the Environment: Pearson New International Edition: An Earth Resource Perspective, 620 p.
- Manolakis, D. G.; Lockwood, R.B. et al. (2016): Hyperspectral Imaging Remote Sensing: Physics, Sensors, and Algorithms, 706 p. Rock, B. N., Vogelmann, J. E., Williams, D. L., Vogelmann, A. F., Hoshizaki, T., (1986): Remote detection of forest damage: Bioscience, v. 36, p. 439-445.
- Sabins, F. F. (1996): Remote sensing: Principles and interpretation. - 3rd ed., 450 p., Freeman Press (San Francisco)

## **Part D**

# **Cartography and Mapping**

Dr.-Ing. Annette Hey, Prof. Dr.-Ing. Ralf Bill and M. Eng. David Hennecke





# 1 Introduction

## 1.1 What is cartography?

The object of cartography is to present the real world in maps and map-related presentations. A map is produced based on the landscape (Figure 76). But what defines a “map”?

**Definition:** A map is a down-scaled, simplified (generalised), in its contents supplemented, elucidated, analogue or digital ground view of the Earth (or of parts of the Earth) or other terrestrial globes and the space projected into a plane.

The first step to produce a map is to transform the image of the real world, e.g. an aerial photograph, into the plane of the map. A map-projection is therefore defined. **Map-projections** have different attributes. Some projections are isogonic (conformal), which means that shapes are not distorted in the map. Other projections are equal to areas or distances. No matter which one is chosen, there is no map-projection that is free of distortions. The reason for this is that the Earth is not a plane, it is a curved figure approximated by an ellipsoid or a sphere, which cannot be projected on a planar sheet of paper without any distortions. After the projection into the plane, the map coordinates are added so that positioning is possible (for more detail on coordinate systems see Part B of this textbook). Because the presented area cannot be depicted in original size a **scale** is needed. “Round” scales such as 1:50,000 or 1:10,000 are preferred.



Figure 76: From reality to a map

A map is an abstract image of reality – objects in the real world are translated into objects on the map (creating a so-called **cartographic model**). Not all objects from the reality will be included in the map. A selection is made as to which objects are relevant. Similar objects are grouped into object classes. Because an unlettered map transmits only relatively little information, labels and a **map legend** (map key) are added.

In general cartography can be described as “the art or technique of making maps or charts” (The American Heritage Dictionary of the English Language, 2007). Besides the technical rules to produce a map there is also the artistic aspect of designing a map, which has unfortunately suffered a decline in its importance due to growing automation. There are definitions of the term ‘cartography’ which are more extensive. Here are two of them:

- “A field that deals with the collection (input), storage, processing and interpretation (analysis) of geospatial data and especially with the visualisation by a cartographic (two-dimensional exemplary) display (presentation).”(Hake and Grünreich, 1994)
- “Cartography is the organisation and communication of geospatial data in a graphic (analogue) or digital way. The cartographic process covers data capturing, cartographic processing and visualising to map use.” (Taylor, 1981)

## 1.2 Map use

Map use is distinguished into two types: active and passive map use. The aim of cartographers today is to support **active map use**. By involving the map user actively in data exploration and presentation, much more information can be communicated and received. Active map use requires a high amount of interactivity and data access. The classical (analogue) paper map allows only a predominantly **passive map use**.

A special type of map use is cartometry. **Cartometry** is the measuring of geometric values in maps, e.g. coordinates, directions, distances, areas, elevation and slope of the terrain. There are different tools for cartometry in analogue maps, e.g. „Planzeiger“ (coordinates and distances), divider (distances), planimeter (areas). While **map-reading** is ‘pure decoding’, such as estimating the number of objects and comparing them, **map-interpretation** is the interpretation of the spatial references of visualised data by using pre-suppositional knowledge of the interpreter. Map-interpretation focuses on thematic aspects.

But what are maps for? The main functions of maps are:

- Decision-making.
- Awareness and education.
- Orientation and navigation.

## 2 First steps to create a map

### 2.1 Map scale

The map scale is given as ratio 1:scale denominator. It determines the ratio of downsizing between the ground view component of a distance in reality and the distance in the map. This ‘strict proportionality’ is only valid if there is little or no generalisation. Especially in small-scale maps, the scale is not equal in the whole depicted area.

**Check:** The **larger** the **scale denominator** the *smaller* the *scale* of the map!

There are two types of map scale – the **numeric map scale** (e.g. 1: 5000) and the **graphic map scale** (e.g. a scale bar). Numeric map scales are sometimes supplemented with examples like ‘1 cm in the map corresponds to 2 km on the ground’. Graphic map scales are useful for simply estimating distances or slopes without calculating. They are important if the map will be copied and thus is enlarged or reduced. While a numeric map scale will become invalid during enlargement or reduction, a graphic map scale will remain valid.

The choice which map scale shall be used depends on the aim of the map. Some general information as to which map scales are often used for which applications is shown in Table 7.

Scale	Applications
large scale (1:500 to 1: 5,000)	cadastre and real estate mapping utility mapping (gas, power, water supply) urban planning in detail
medium scale (1: 10,000 to 1: 250,000)	topographic mapping thematic mapping <ul style="list-style-type: none"> <li>• urban planning (zoning maps)</li> <li>• regional soil maps</li> <li>• regional geological maps</li> <li>• regional biotope maps</li> <li>• ...</li> </ul>
small scale (1: 500,000 and smaller)	thematic maps (national/global coverage) atlases

Table 7: Map scales and typical applications

Map scales are distinguished between original scale (scale of survey) and secondary scale. Secondary maps are derived from a base map (original scale) by generalisation. In transferring a map from one larger scale into a

smaller scale (because of a smaller available display area) some steps of simplification and reduction are necessary. This process is referred to as 'generalisation'.

## 2.2 Generalisation

To display the real world in a map, a model of reality must first be created. An important part of modelling is to simplify because the real world is much more complex than can be presented in a map. One reason for this is the limited space in a map. Another reason is the demand for legibility and the individual aim of the map. The result is that not all data and information from reality can be presented in the map. The same problem occurs when transferring a large scale map (original map) into a small scale map (secondary map). To solve the problem an abstraction has to be performed.

**Definition:** Cartographic generalisation is a collection of theories, methods and procedures to reduce and generalise cartographic information. In this process, certain subsets are selected from an information set and summarised in superior units. The transformed pieces of information are adapted due to the scale or reduced map surface, or they are simplified or visualised according to the purpose of the map.

In the procedure of transferring object data into a basis map or transferring a basis map into a secondary map it is important to depict reality in its most important and typical attributes and its characteristics according to

- the purpose,
- the theme and
- the scale

of the resulting map.

There are two types of generalisation:

- **Survey generalisation** or generalisation in data capturing: conversion from reality to basis map (primary data capture).
- **Cartographic generalisation:** conversion from a basis map into a secondary map with a smaller scale.






















	generalisation step	cartographic visualisation		
		in the <b>base map</b> e.g. 1 : 10 000	in the <b>secondary map</b> e.g. 1 : 50 000	re-enlargement to original scale
geometric	1. simplification			
	2. enlargement			
	3. displacement			
factual with geometric effect	4. subsuming			
	5. selection			
	6. stereotyping			
	7. evaluating			

Figure 77: Steps of cartographic generalisation (after Hake et al., 2002)

Survey generalisation is the conversion of real world objects into abstract map objects. It is the first step of cartographic modelling. Objects are selected and grouped according to conceptual community. Geometric shapes are simplified.

Besides a factual (conceptual) simplification and selection of objects a cartographic generalisation is also a graphical simplification and a generalisation regarding geometry (position). The process of cartographic generalisation can be structured into seven steps:

- simplification
- enlargement
- displacement
- subsuming
- selection
- stereotyping
- evaluation

with Figure 77 illustrating these steps.

Mathematical formulas for describing the process of generalisation have been developed from empirical research. With the following equation, the length of linear map objects in the secondary map can be estimated according to the length in the base map.

$$D_S = D_B \times \sqrt{\frac{S_B}{S_S}} \quad \text{where:}$$

$D_S$  : Distance in secondary scale  
 $D_B$  : Distance in basis scale  
 $S_B$  : Scale denominator of base scale  
 $S_S$  : Scale denominator of secondary scale

The ‘Object-Selection-Rule’ can be used to estimate the number of objects in the secondary map according to their number in the base map.

$$n_S = n_B \times \sqrt{\frac{S_B}{S_S}} \quad \text{where:}$$

$n_S$  : Number of objects in secondary scale  
 $n_B$  : Number of objects in base scale  
 $S_B$  : Scale denominator of base scale  
 $S_S$  : Scale denominator of secondary scale

In connection with generalisation, the limits of cartographic presentability have to be considered. There are **recommended minimum sizes of map objects** that should be regarded. These minimum sizes indicate the limit of legibility. Coloured map objects and map objects on a coloured background should be larger than map objects in black and white presentations because the contrast is weaker. Some recommended minimum sizes (for paper maps) are:

- stroke width, b/w: 0.05 – 0.08mm
- stroke width, coloured: 0.08 – 0.10mm
- interspace between strokes or areas: 0.15 - 0.25mm
- quadrangle, filled: 0.40×0.40mm
- rectangle, filled: 0.0×0.60mm
- point diameter: 0.25mm
- labels: 6pt

For digital maps (displayed on a screen), the minimum sizes are different. Because of the pixel structure and the lower resolution compared to paper maps a magnification factor of between 2 and 4 has to be applied to the recommended minimum sizes for paper maps.

There are two different methods to generalise areas; the selective method and the individual method (Figure 78). While the **selective area generalisation** works with minimum sizes and can therefore be quite easily automated (and is largely objective), the **individual area generalisation** is very subjective. The advantage of the individual method is the preservation of the natural area proportion.

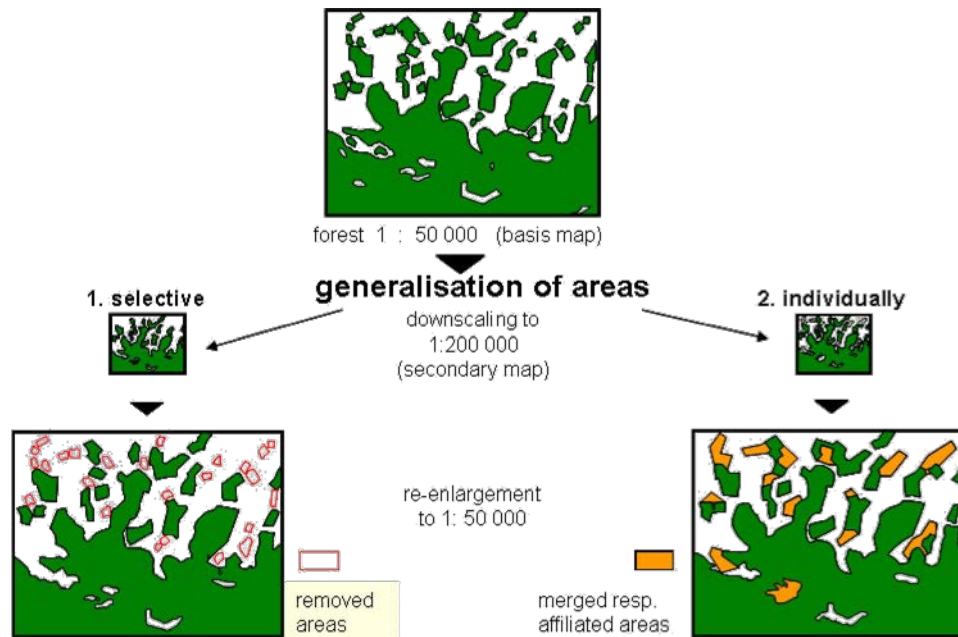


Figure 78: Generalisation of areas

### 3 Symbolisation

#### 3.1 Graphic variables

To visualise object attributes in thematic maps, the graphical variables of the map symbols are varied (Figure 79). According to Bertin (1983) the graphic variables are:

- Colour
- Shape
- Size
- Brightness
- Texture
- Orientation

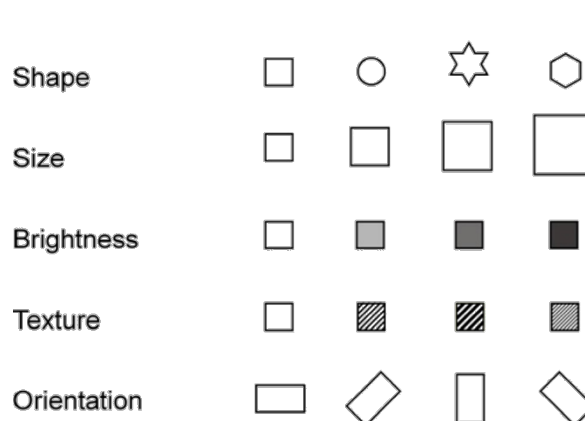


Figure 79: Possible variation of graphic variables

The variation of these graphic variables causes different effects that are used to display attributes of map objects in the map.

The variable colour is selective (qualitative).



The variable brightness is assortative.



The variable size is the only one that is quantitative.



The length of a graphic variable indicates how many values can be represented legibly and distinguishably by varying this variable. For the variables size, brightness, and colour the following lengths are generally estimated:

- Size: approx. 5
- Brightness: approx. 7
- Colour: approx. 8

### 3.2 Symbol scale

For the alteration of the variable size, a symbol scale is defined. This symbol scale is largely independent from the map scale. There are three types of symbol scales:

- continuous
- graduated
- unique value symbols

Fundamentally there is a choice between length-, area- and volume-proportional symbol scales (Figure 80, Figure 81 and Figure 82). There are also intermediary symbol scales which are based on mathematic formulas. They combine the higher degree of being correctly estimated, as with area-proportional symbol scales, with the greater value range that can be represented by using a volume-proportional symbol.

Choosing an appropriate symbol scale is an optimisation problem. There are various aspects to be considered. On the one hand it is necessary to consider which demands are made regarding the correct estimation of presented values, on the other hand the value range that must be presented, the space available in the map and the decision whether and how the data should be classified (individual values or grouped values) affect the choice of symbol scale. In most cases not all requirements will be satisfied at the same time. Empirical studies have shown that length-proportional symbol scales are estimated much better than area- or volume-proportional symbol scales. The error in estimation increases significantly with the latter two scale types. The value margin that can be presented depends on the kind of symbol scale. Length-proportional symbol scales can only present a value range of about 1:100 (ratio minimum:maximum). Area-proportional symbol scales allow value ranges up to 1:1000. Above this, a strict area-proportional scale is not possible. The space available in the map is an important influence factor when choosing a symbol scale. Length-proportional figures grow much faster than area- or volume-proportional ones. The degree of quantitative generalisation (classification) is determined by the number of classes. A presentation of individual values (no classification) is also possible.

The definition of a symbol scale starts with the definition of basis values, e.g. 1cm corresponds to 500 sheep (length-proportional) or 1cm<sup>3</sup> corresponds to 200 tonnes of potatoes (volume-proportional). According to this base value, all other values can be transformed into the according variable (length, area or volume).

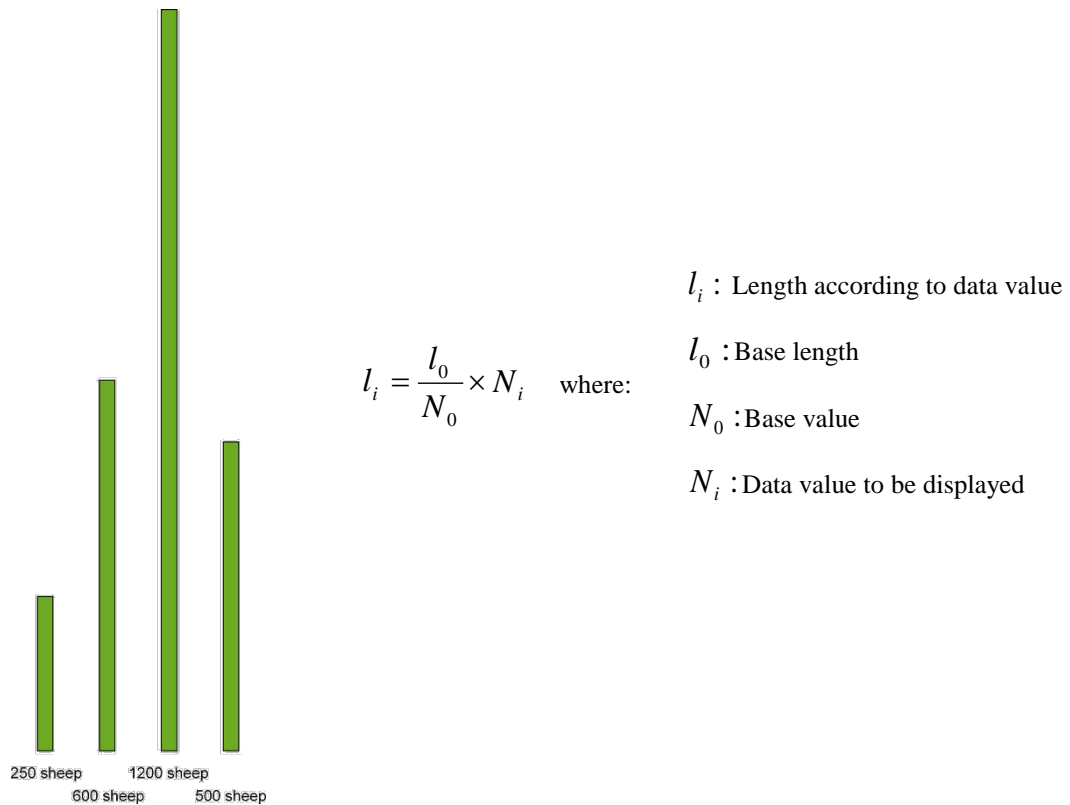


Figure 80: Length-proportional symbol scales

Area-proportional symbol scales can be applied to different geometric shapes. These shapes should be symmetric. Pictographic symbols should not be used with symbol scales. A very simple (and therefore particularly appropriate) geometric shape is the rectangle (Figure 81).

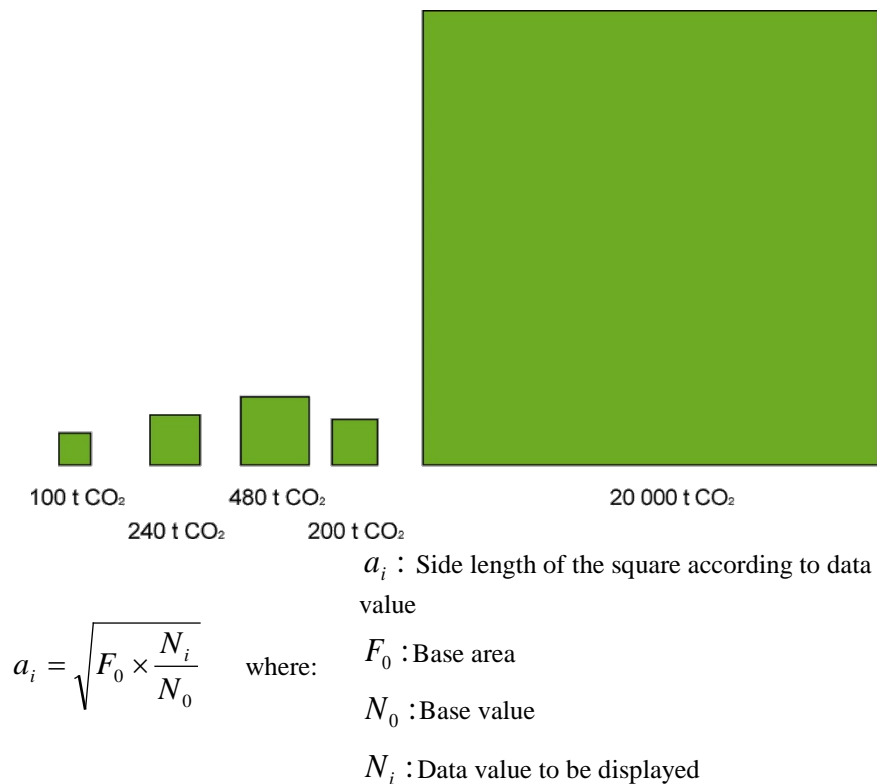


Figure 81: Area-proportional symbol scale



As well as area-proportional symbol scales, volume-proportional scales can also be applied to different geometric shapes. With volume-proportional symbol scales the symmetry of the geometric shape increases in importance because the presentation of the (three-dimensional) figure in the map will be only two-dimensional. This means that some parts of the figure will be invisible and have to be completed in the mind before the presented value can be estimated. The example given in Figure 82 is a cube.

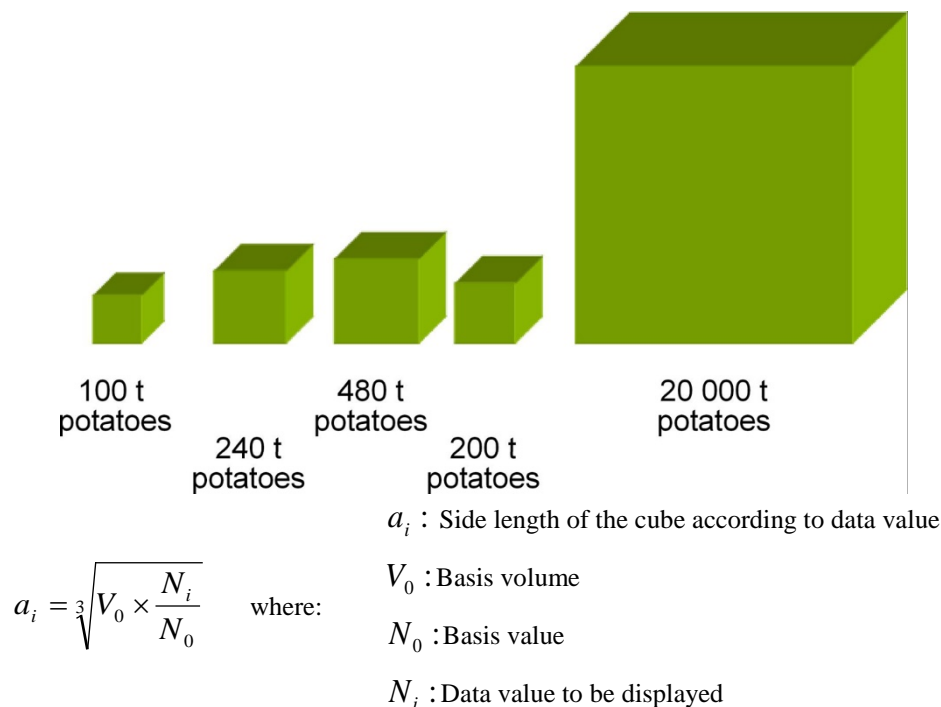


Figure 82: Volume-proportional symbol scale

### 3.3 Presentation of relief and terrain

The presentation of the third dimension in a two-dimensional medium (map) is difficult. In historic times, artistic methods were used often to visualise relief and terrain. These maps were vivid but a rule-based production was impossible. Today's maps use abstract methods as well as vivid techniques (hill-shading). Examples of these abstract methods are **contour lines**, contour-level-colouring and points with given height information, e.g. trigonometric points.

**Definition:** Contour lines are connecting lines between points of the same height relating to a defined ground level that are projected orthogonally into the plane of the map. A contour line is also referred to as 'isohypsis'. (Likewise a depth contour is called 'isobath'.)

The colour of contour lines is often adapted to the type of relief in the area. Contour lines in rocky areas are black, while contour lines on glaciers are blue. Contour lines in other areas are brown. Labels of contour lines are always placed with the base line towards the valley. This supports relief recognition. The map users can imagine themselves standing at the bottom of a valley looking up the hills and being able to read the height of each contour line. Thus the question 'which way is up?' can be answered quite easily.

To distinguish positive forms (e.g. hilltops) from negative forms (e.g. depressions), a special method of presentation is used (Figure 83). With this, a clear identification is possible.

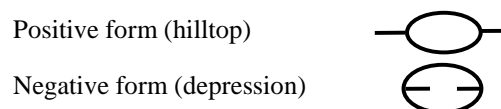


Figure 83: Contour lines for positive and negative relief forms

An important parameter of contour lines is the **equidistance**.

**Definition:** The equidistance tells the constant vertical distance between adjacent contour lines (= height interval). The smaller the equidistance the closer the landscape is described.

With a small equidistance, the contour lines may appear too close together such that the map is hard to read. The choice of which equidistance should be used has to be made in consideration of the type of relief. One basic rule is: smooth forms, such as hilltops and basins, are more important in describing the landscape than sharp forms, like rock pinnacles. If the landscape contains steep passages (large equidistance) as well as flat ones (small equidistance), intermediate contour lines can be applied. The purpose of these intermediate contour lines is to give some information as to the shape of the land between the regular contour lines, which is especially important in flat regions where the distance between adjacent regular contour lines can be very large.

Errors in the contour lines lead to misinterpretation of relief behaviour. Errors in positioning the contour lines in the map will cause errors in determining the height of points. If the height of a contour line is wrong, the correct position of a point with known height cannot be found. Coarse errors will be visible in the map as peaks and troughs.

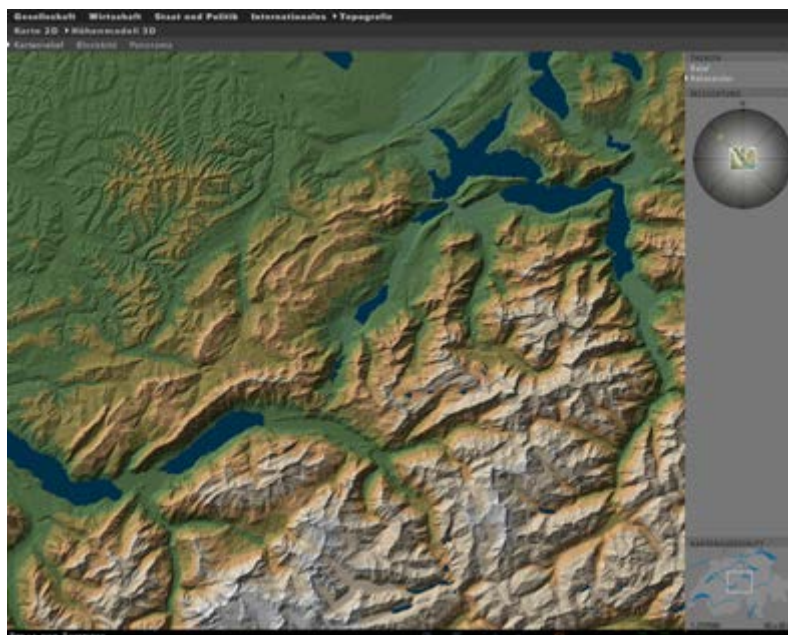


Figure 84: Contour-level-colouring with hill-shading (Source: Atlas Schweiz)

Another abstract method to visualise the relief is contour-level-colouring (Figure 84).

**Definition:** Contour-level-colouring is a method of presenting the relief by colouring the interspace between certain contour lines according to a given colour ramp. This method is also referred to as ‘hypsometric colouring’ or ‘altitude tint’.

There are different approaches which colour ramp should be applied. Some of the most common are shown in Table 8.

Year	Author	Approach
1847	E von Sydow	The higher, the brighter
1867	F. Hauslab	The higher, the darker
1930	K. Peucker	Spectral adaptive colour ramp from grey (low) to red (high): see Figure 85

Table 8: Approaches for defining a colour ramp



Figure 85: Colour ramp according to Peucker

A method of presenting the relief that is adapted to the visual experience is hill-shading. Based on an assumed incidence of light, shadows are added to the relief presentation to create a three-dimensional impression. The direction of the assumed incidence of light is most often north-west, even if the sun only seldom appears there. However, if the direction is changed to south-east, an effect called ‘relief reversion’ occurs: valleys appear to be hilltops while mountain ranges appear as valleys (Figure 86). This effect can be resolved by selective mental counter-steering, but this demands some concentration. A map should be read as spontaneously as possible, therefore an illumination direction from south or southeast should be avoided. A possible explanation why we expect light to shine from north-west is that an artificial light source (candle or lamp) is placed left above a sheet of paper when writing with the right hand. Throughout the last centuries this impression may have become persistent and so has found application in cartography. Although this is a rather weak explanation, other explanations have not yet been found. Today satellite images (where illumination direction is often not north-west) are first de-shaded, which means all shadows are removed, and afterwards a shading with an assumed illumination direction from north-west is added.

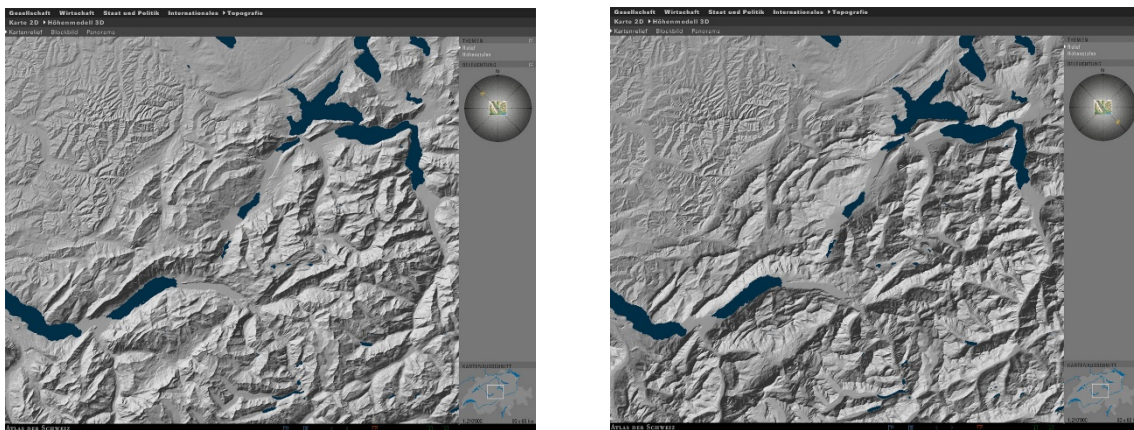


Figure 86: Hill shading with illumination direction (a) north-west and (b) south-east (source: Atlas Schweiz)

## 4 Labelling

### 4.1 Map fonts

Map fonts should be legible even at small sizes. Sans-serif fonts, such as Arial or Verdana, should therefore be used. The font size should never be below 6pt. To ensure an explicit assignment, the labels should be placed as close as possible to the labelled object. A ranking of label positions in relation to the relevant object (Figure 87) has been evolved out of experience:

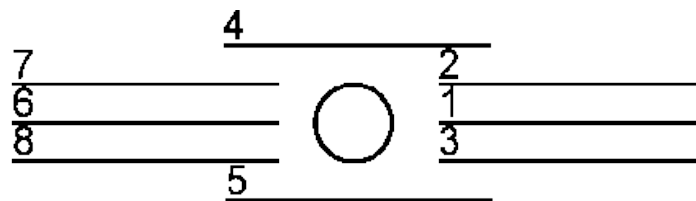


Figure 87: Label position ranking

The whole lettering of the map should be in reading direction from the southern map frame (see Figure 88). The offset between label and related object should be well-balanced. When using a small font size the offset should also be small, large font sizes require a larger offset. Overlapping labels should not appear in a map. Names should never be separated, except where there is an acute shortage of space. Syllables should never be separated.

There are some basic rules for the alignment of labels. The font baseline should be parallel to the upper or lower map frame or parallel to the next circle of latitude when using a geographic coordinate system. The labels of settlements near large water bodies are subject to a special rule. If the settlement is located directly on the coast, the label **can** be placed in the water surface. If the settlement is located inland then the label **must** be completely within the land area.

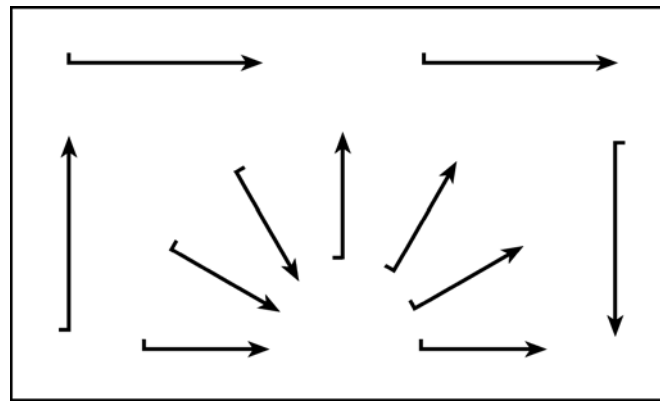


Figure 88: Reading direction of map labels

When labelling linear objects, the label should be parallel to the linear object (Figure 89a). A horizontal or near-horizontal label position is to be preferred. Area-related labels should be adapted to the main extent direction (if existent). The label may be spaced, but still has to be legible continuously (Figure 89b).

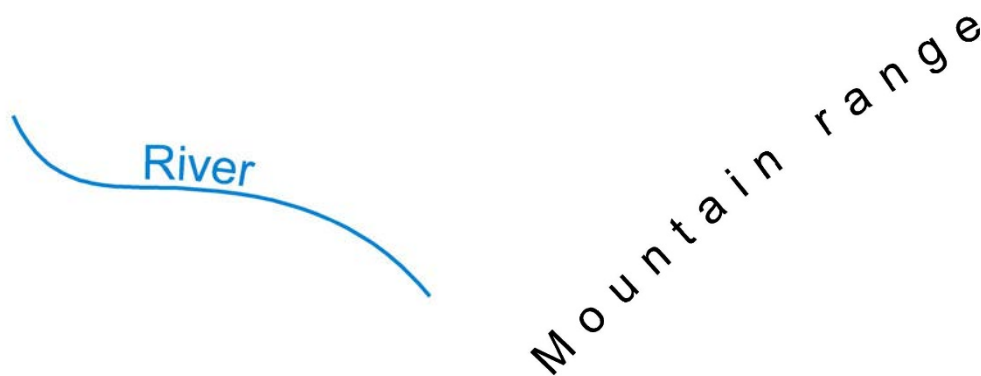


Figure 89: Label placement on (a) linear and (b) area-related objects

The spelling of geographic names requires special attention, especially in areas outside the cartographer's own language area. If Latin letters are used in these areas, a transformation can be made without any difficulty by using so-called diacritical signs (accents). To transform non-Latin scripts, one of two methods – transliteration and transcription – can be applied.

**Definition:** Transliteration is the letter-wise transformation from the orthography of the original language to the orthography of the target language. In the process the pronunciation is not considered.

The advantages of transliteration are that there are already standard transliteration alphabets (ISO) and a re-transformation is possible, because each letter is clearly defined. Disadvantages are the need for additional diacritical signs and the sometimes unusual letter combinations. The transliteration method is not applicable to phonetic spellings and pictographic languages (e.g. Japanese).

**Definition:** Transcription is the transformation of names based on the pronunciation in the original language into the orthography of the target language. In this process no diacritical signs are used (e.g. č → tsch).

Advantages of the transcription method are that the transcribed names can be read very easily and that the correct pronunciation is given. Disadvantages are that a re-transformation is impossible and that the transcribed names need much more space in the map than the original ones.

## 4.2 Legend design

**Definition:** The legend or symbol key is the summarising explanation of all graphic elements used in the map. It is the key to understanding (decoding) the map content.

Basic requirements for a legend are:

- Completeness (explain **all** map symbols!).
- Correctness (same size of map symbols as in the map!).
- Clarity and intelligibility (logical composition, arrangement in groups).

The design (layout) of a legend should follow certain rules. **Point map symbols** should be arranged on an axis. In connection with the use of symbol scales, the requirement of correctness here is especially important. The best way to explain diagram figures is to use an example diagram. When explaining **linear map symbols**, all symbols in the legend should be of the same length. A difference in length implies differences in importance, which are most often non-existent. Structured lines (e.g. dashed-dotted, dashed) should always be explained using at least 3 parts. There are some special requirements when explaining arrows. Preferably they should be horizontal and pointing to the greatest space (map surface). If the arrows stand for opposing statements, they should also point in opposing directions in the legend. Map symbols which refer to an area (e.g. fill colours, hatching and textures) are often explained by using rectangular boxes. These boxes should have an outline. Never use shadows to give the boxes a three-dimensional appearance. Continuous series of boxes (no gaps) should only be used if it is a spatially continuously variable quantity that is represented (e.g. rainfall). If it is a spatially discontinuous quantity (e.g. population density, which will normally be calculated based on administrative regions), separate boxes should be used to represent it. The position of high and low values is as follows: high values on top or to the right; height- and depth-levels according to natural order. Qualitative data should always be explained by using separate boxes. Areas with individual outlines (e.g. surface mining) do not have to be explained with boxes.

If font size, font colour, font inclination (regular, italic) or the typeface have a meaning, they have to be explained. Names should be summarised in groups according to themes. They should be positioned left-aligned and they should be of the same length. If there is a ranking (e.g. capital, district town, rural community) they should be arranged top-down.

In general, the **primary rank order** of map symbols in the legend is according to the content starting with point map symbols, followed by linear map symbols and finally map symbols referring to an area. Two map symbols which **factually belong together** (e.g. castle and castle ruin), can be arranged side by side.

There are two types of legends. According to the extent of the map there are map surface legend and map frame legends (Figure 90).

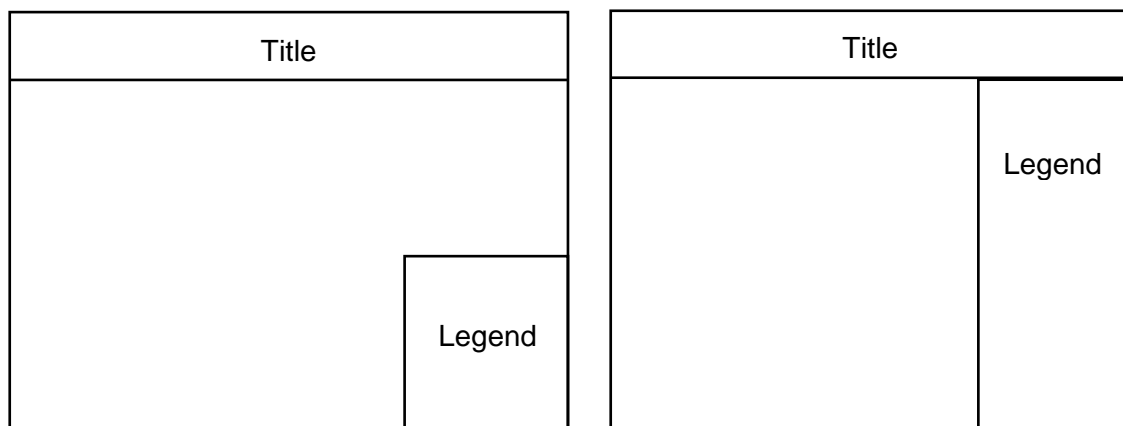


Figure 90: Map legend types – (a) map surface and (b) map frame

## 4.3 Map layout

The layout of a map consists of the **title (theme)** and if necessary a subtitle, the legend (symbol key), the map scale (numeric, if possible additionally graphic), reference system (coordinates), method of data capturing respectively basis map, the date of data capturing respectively the date of cartographic editing, data sources and licences, the author and if necessary the editor and the publisher (appendix maps). A **north-arrow** is only needed if the map is **not pointing north** or if the north direction cannot be determined from the map (e.g. if there is no coordinate grid given). There are two types of map layout: frame-map and island-map (Figure 91). A **frame-map** is a map whose map surface is bordered by the map frame (Bollmann and Koch, 2002). An **island-map** is a map

whose map surface does not reach until the map frame. The presentation is limited to an irregular shaped area, which is most often given by administrative borders (Bollmann and Koch, 2001)

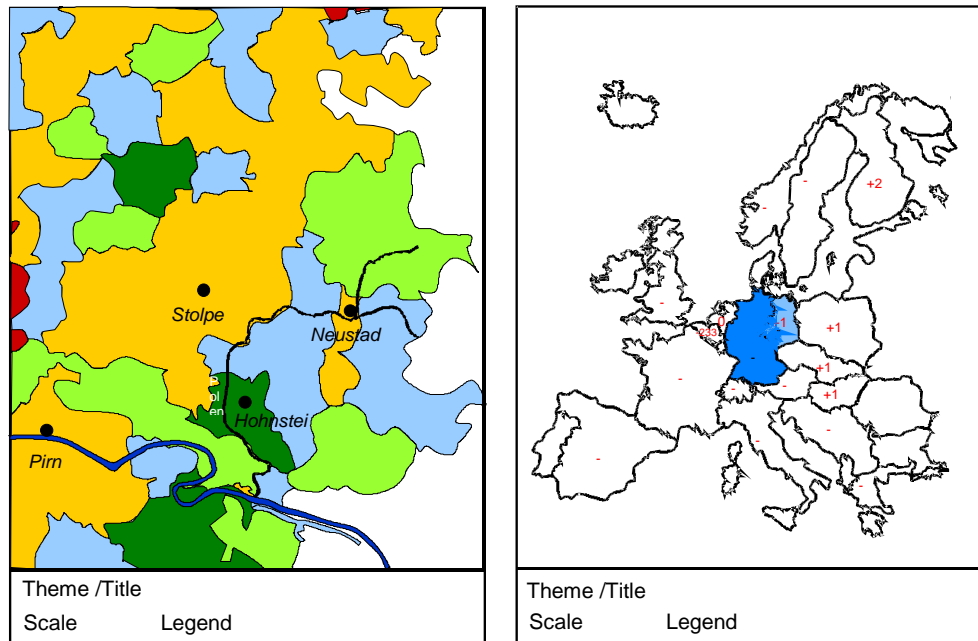


Figure 91: Map layouts – (a) frame map and (b) island map

## 5 Topographic maps

### 5.1 Introduction

**Definition:** A topographic map is a down-scaled, simplified (generalised), in its contents supplemented, elucidated, analogue or digital **ground view of the Earth** (or of parts of the Earth) or other terrestrial globes and the space **projected in a plane**.

The main function of a topographic map is navigation and orientation. In the map surface, the so-called cartographic situation is presented. The parts of the cartographic situation are shown in Figure 92.

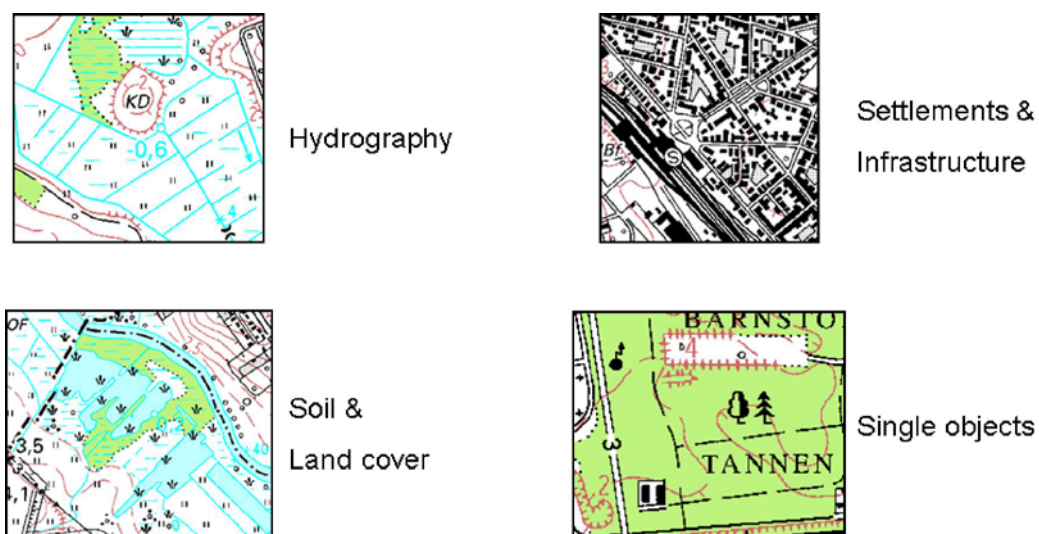


Figure 92: Elements of the cartographic situation

A **topographic map** differs from a thematic map in that a thematic map does not present the geographic space itself (the real world) but concrete and abstract spatial **phenomena** (circumstances and situations) and **processes**

both of the natural and of the socio-economic field of geographic space. A **thematic map** always relies on a topographic basis. Next two figures illustrate the difference between topographic and thematic maps.



Figure 93: Topographic map (source: Atlas Schweiz)

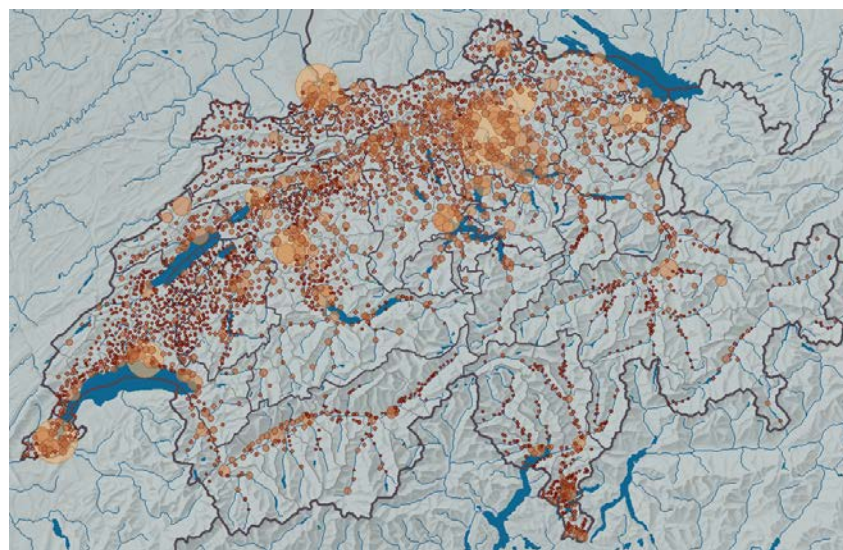


Figure 94: Thematic map of employees (source: Atlas Schweiz)

## 5.2 Comparing topographic map series in two countries

The production and updating of the official topographic maps is the responsibility of national mapping and surveying organisations. Typical for official topographic maps are map series.

### Germany

In Germany topographic maps in the scales 1:5000 (western part), 1:10 000 (eastern part), 1:25 000, 1:50 000, 1:100 000 and 1:200 000 are available. The small scales are derived from the topographic base maps (1:5000 respectively 1:10 000) by using generalisation. A part of this map series and the effects of generalisation is illustrated (Figure 95).

### Vietnam

The national mapping agency of Vietnam is the Division of the General Directorate of Land Administration (Tong Cuc Dia Chin (TCDC)) in Hanoi. Maps are published by the Cartographic Publishing House. Map sales are organised by the General Directorate's Center for Research of Land Administration. To cover the whole area of Vietnam a different number of map sheets is needed according to the map scale. The figures for Vietnam are:

1: 50 000	588 sheets needed
1: 100 000	164 sheets needed
1: 250 000	40 sheets needed

There are larger scale photogrammetric series, e.g. in the scales of 1:10 000 and 1:25 000, that do not cover the whole country. (Parry, & Perkins, 2002). In 1996 the first National Atlas of Vietnam was published. Figure 96 shows examples of Vietnamese topographic maps.

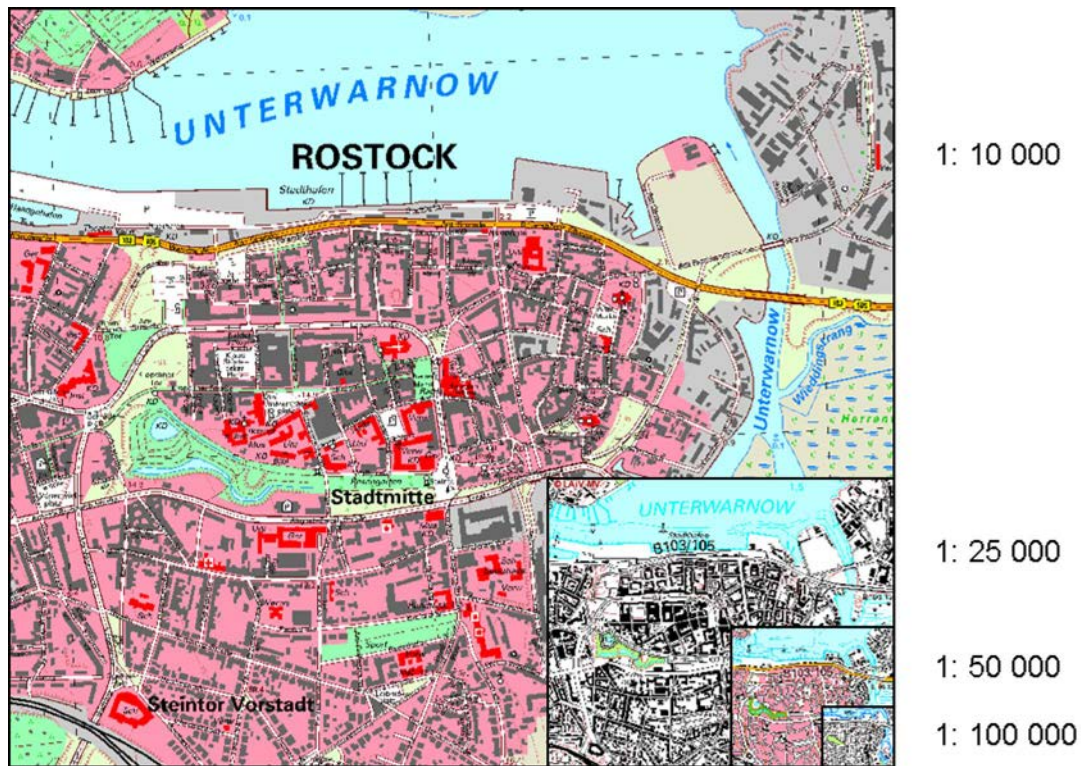


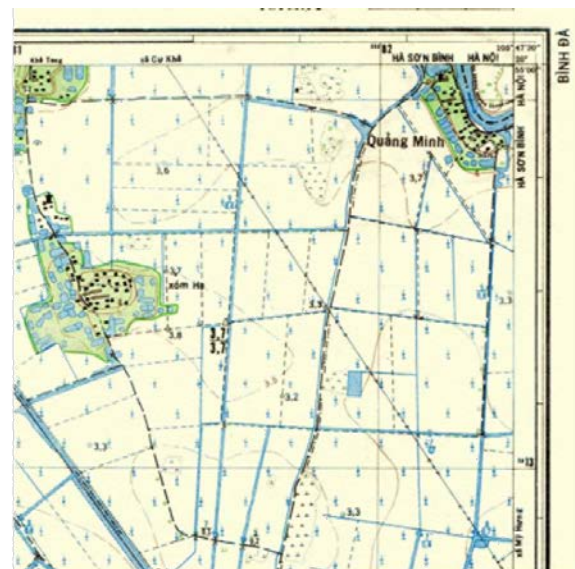
Figure 95: Topographic map series (Germany)



1: 50 000



1: 25 000



1: 10 000 (Hanoi)

Source (image): [www.omnimap.com](http://www.omnimap.com), July 2007

Figure 96: Topographic maps in Vietnam



## 6 Thematic maps

### 6.1 Introduction

A thematic map goes beyond the representation of topography.

**Definition:** A map to visualise concrete and abstract, spatial **phenomena** (circumstances and situations) and / or **processes** both of the natural and of the socio-economic field of geographic space (always relying on a topographic basis).

The structure of a thematic map consists of a topographic basis, the special thematic content and the map frame designed according to the purpose of the map.

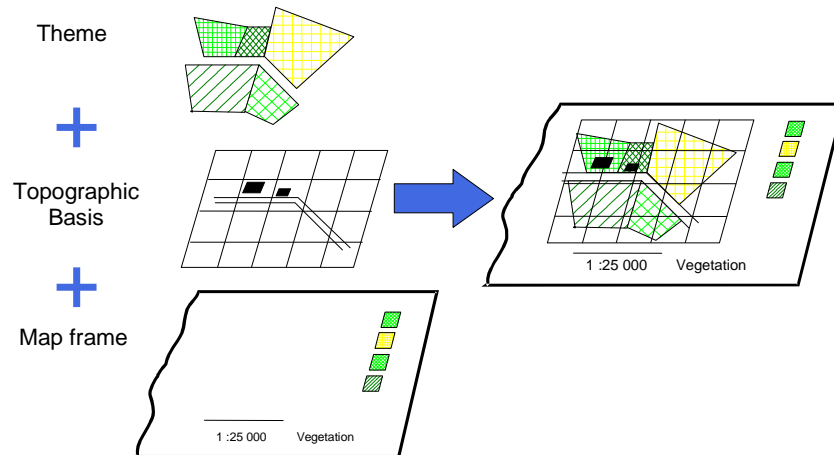


Figure 97: Structure of thematic maps

### 6.2 Types of data

There are some general rules that have to be taken note of when designing thematic maps. Essential and intense objects should be emphasised by using intense colours or line weights. To represent similar or related objects, the same point or linear symbols should be used. Equal colours, textures or hatches of areas can be used as well. If a fact is uncertain or controversial (e.g. the course of a border), this attribute should be represented by using dotted or dashed line symbols, colour gradients or interlocking of areas, special signatures, colours, textures or hatchings. The special circumstances can be indicated by giving additional information (e.g. question marks). Differences in quality are visualised with different point or linear symbols as well as different colours, textures or hatching of the areas. Differences in quantity should be indicated by using different sizes (point and linear symbols: length, height and width) or different intensities of colours, textures, or hatching of the areas. If there is a thematic correlation between objects or circumstances, the signatures should be grouped accordingly (e.g. all objects concerning water are represented in blue). The most important rule for designing maps and map-related representations is that the cartographic visualisation cannot be more exact than the primary data used to create it.

To apply these rules correctly, the different data types and their specific attributes must first be understood. Data types can be coarsely classified into qualitative and quantitative data. **Qualitative data** are distinguished into nominal scaled data and ordinal scaled data. **Nominal scaled data** only give information as to whether two objects are similar or dissimilar. Scale limits are not defined. An example of nominal scaled data is: river, building, road (Figure 98).



Figure 98: Qualitative scaled data

**Ordinal scaled data** start at a defined scale starting point. The scale is open-ended. Ordinal scaled data determine a ranking. A typical example of ordinal scaled data is the ranking of borders (e.g. national border, state border, district border) as in Figure 99.

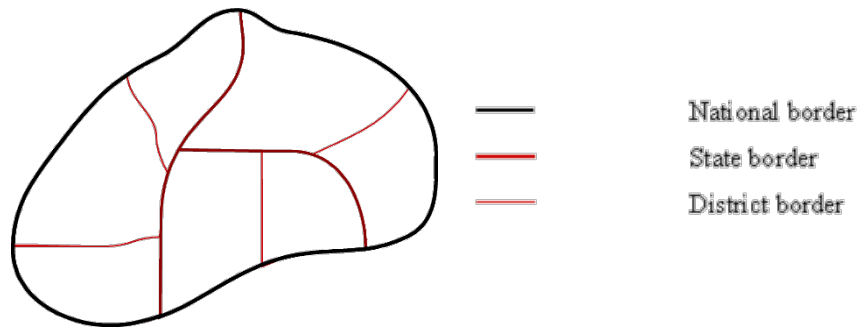


Figure 99: Ordinal scaled data

**Quantitative data** is distinguished into interval scaled data and ratio scaled data. The scale of **interval scaled data** has no start and no end; only the point of origin is defined. Interval scaled data allow additive operations (addition and subtraction). An example of interval scaled data is temperature data in degrees Celsius.

The scale of **ratio scaled data** starts with an absolute zero point and is open-ended. All kinds of mathematical operations (additive and multiplicative operations) can be applied to ratio scaled data. Ratio scaled data are for example harvest amounts. For these data e.g. sums, differences or a ratio of amounts can be calculated.

Another important attribute of quantitative data is the reference. Quantitative data without any reference is called 'absolute values'. **Absolute values** are for example a number of objects such as 500 sheep. Quantitative data with a reference can refer to an area or a basic amount. Data referring to an area are quite common among spatial data. A typical example is the population density (inhabitants per km<sup>2</sup>). Data referring to a basic amount require another representation method (e.g. coloured signatures with an applied symbol scale representing the base amount) than data referring to an area (reference area is coloured). An example of quantitative data referring to a basis amount is the number of students referring to the total number of inhabitants.

Quantitative data is often generalised before it is visualised. This quantitative generalisation is referred to as **classification**. A classification shall enable a reasonable representation, which is impossible when the number of different data values is higher than approximately 15. The figure below shows a representation of population density, where each value is assigned a unique colour (due to the high number of different data values the colours are not specifiable).

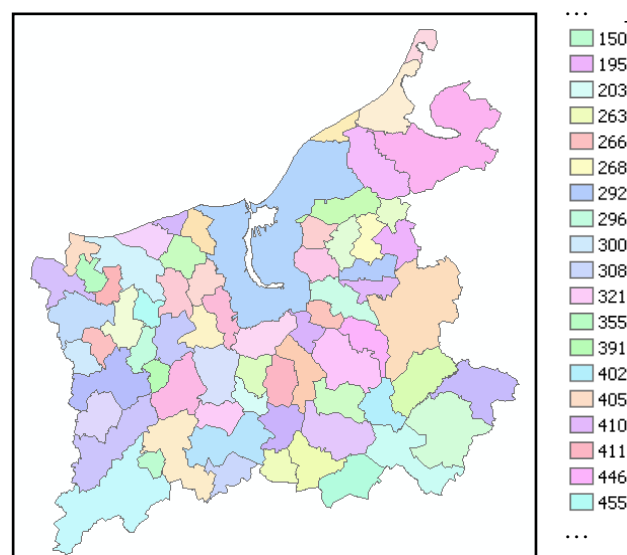


Figure 100: Visualisation without classification

### 6.3 Classification methods

The purpose of a classification is to allow the creation of clearly-arranged maps and avoid high signature pressure. It should be visible on first sight whether features belong together or not.

**Definition:** Classifying data is a quantitative generalisation to visualise statistical single values whose distinctive distribution should remain recognisable.

During a classification, different values are merged into **classes**. One principle of classifying data correctly is the demand for completeness. All data values should be included in the classes, which should form a continuous sequence. There should be no gaps between adjacent classes, however the need for a continuous sequence cannot be fulfilled for all kinds of data: if a gap in the domain is a characteristic attribute of the data then a gap between adjacent classes is allowed and required. Another principle of classification is the definition of explicit class limits. The assignment of data values to a class has to be clear. This is achieved by class limits such as 5 to <10, 10 to <15, etc. The number of classes depends on the source data. If the domain of source values is large, more classes will need to be defined than if the domain is small. The accuracy of class limits should not be more exact than the accuracy of the source data. The loss of information due to the classification should be minimised. Open classes such as < 20 should be avoided, otherwise minimal and maximal values will be ‘concealed’. A suitable classification method should be chosen according to the type and distribution of the source data.

Available classification methods are:

- Factual classification (natural breaks).
- Classification based on meaningful thresholds.
- Classification based on mathematical rules.
- Classification based on statistical values.
- Spatial classification.

The **factual classification**, sometimes referred to as the method of ‘natural breaks’, bases on a statistical frequency analysis of the original data. If the statistical analysis is performed using a histogram, the class limits are located where values are missing or where considerable breaks are visible (Figure 101).

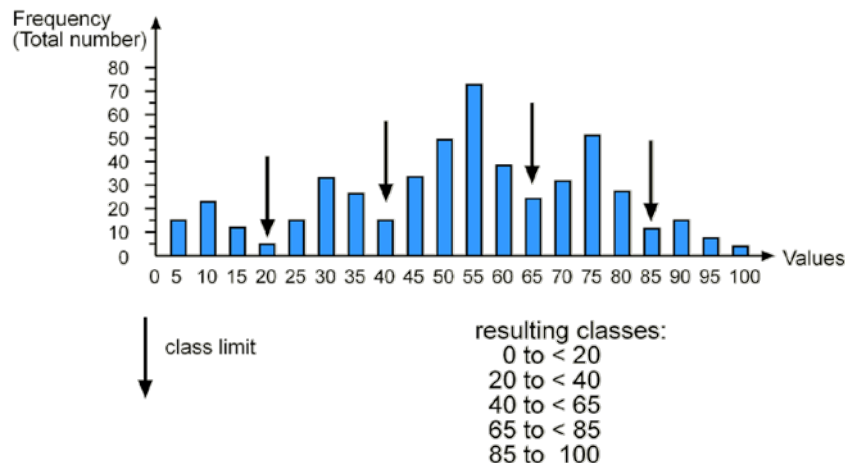


Figure 101: Factual (‘natural breaks’) classification by histogram

If the statistical frequency analysis is performed using a cumulative frequency distribution diagram (sum graph) then the class limits are located at distinctive bends of the graph which indicate changes in frequency (Figure 102)

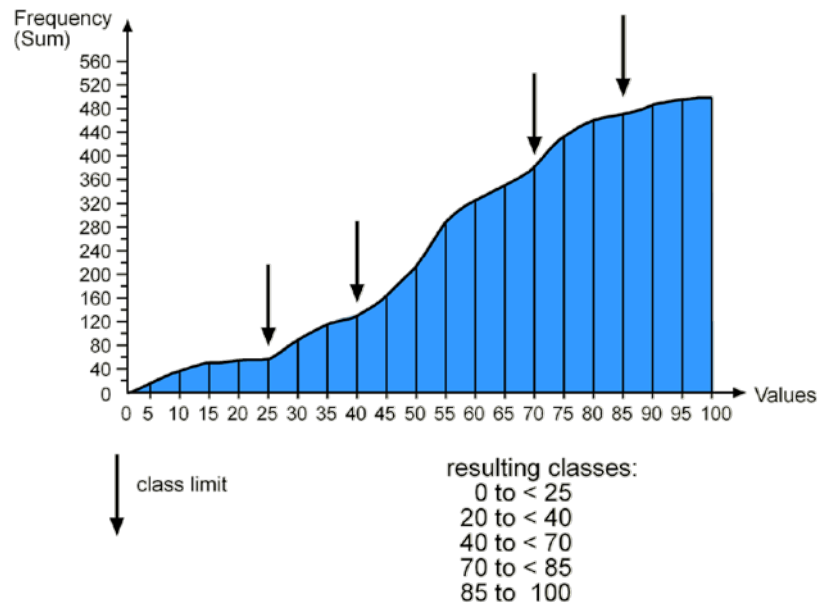


Figure 102: Factual ('natural breaks') classification by cumulative frequency distribution

The classification based on meaning thresholds requires experience: it is defined by experts. A typical example of a classification based on **meaningful thresholds** are slope groups (Table 9).

Slope group	Slope angle	Verbal description
0	0° - 0°59′	plane
1	1° - 1°59′	very flat
2	2° - 4°59′	flat
3	5° - 8°59′	moderately inclined
4	9° - 12°59′	highly inclined
5	13° - 16°59′	steep
6	17° and above	very steep

Table 9: Classification based on meaning thresholds for slopes

The classification based on **mathematical rules** is a mere schematic method. Several mathematic conditions are used as a basis. Equidistant intervals are applied to equally-distributed values. The range between minimum and maximum values is separated into equal intervals  $b$ .  $\rightarrow N_{\min} + b + b + \dots = N_{\max}$

To calculate the interval  $b$ :

$$b = \frac{N_{\max} - N_{\min}}{m}$$

Classification using **arithmetic progression** is applied if a part of the data domain should be more differentiated than the rest. If the lower domain should be more differentiated, class threshold values will increase by a constant amount  $d$  for each class:

$$N_{\min} + d + 2d + 3d + \dots + md = N_{\max}$$

Formula to calculate  $d$ :

$$d = \frac{(N_{\max} - N_{\min}) \times 2}{m \times (m + 1)}$$

The progression can be inverted (upper domain will be more differentiated).

By using the classification method of **geometric series**, the class width increases strongly in the upper domain. An attribute of geometric series is that the ratio  $q$  of two following class limits is constant, mathematically described as:

$$N_{\min} \times q_m = N_{\max}$$

Formula to calculate  $q$ :

$$\log q = \frac{\log N_{\max} - \log N_{\min}}{m}$$

Classification can also be derived from **statistical values**. Often used statistical values are quantile and standard deviation. The quantile is applied to achieve a maximum graphic differentiation. All classes appear in the same number of reference areas in the map. A classification based on the standard deviation illustrates how the larger part of the values is distributed around the average. The class width is the standard deviation starting from the average.

Many visualisation tools already offer some classification methods. ArcMap, the visualisation module of ESRI's ArcGIS, offers the following classification methods (Figure 103):

- Manual input of class limits: This can be used to implement a classification based on meaningful thresholds.
- Equal interval classification: The number of classes is determined by the user. The resulting class limits are calculated by the software.
- Defined interval: The class width is determined by the user. The resulting number of classes and the class limits are calculated by the software.
- Quantile: The class limits are calculated by the software, so that each class contains the same number of objects (values).
- Natural breaks (factual classification): Basing on a statistical frequency analysis of the original data the class limits are set, where a significant decrease in the frequency of values is recognisable or where values are completely missing.
- Standard deviation: The standard deviation or a multiple of it is used as class width. This kind of representation emphasises the distribution of the values around the mean value. Often a bipolar scale (e.g. from red to green) is applied to clearly distinguish values above and below the mean.

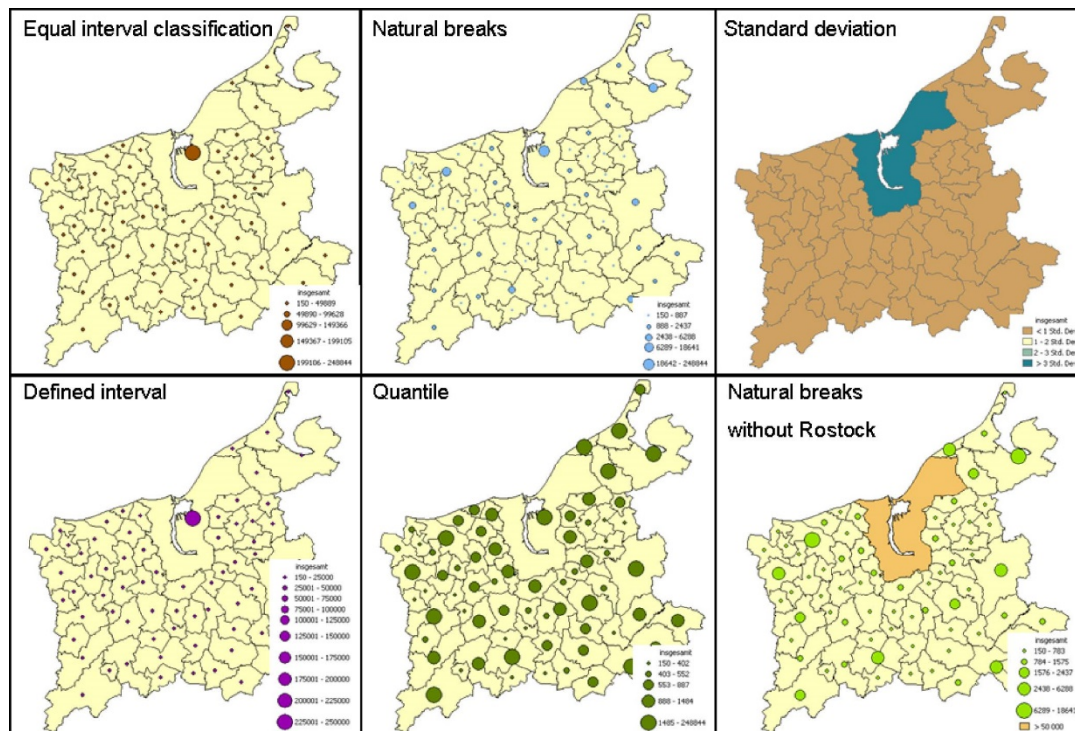


Figure 103: Comparison of classification methods available in ArcGIS

## 6.4 Presentation methods

Presentation methods are always chosen according to the data to be represented. There are some basic rules regarding the eligibility of certain methods. Quantitative data, that means data, that answers the question “how much?”, are visualised differently according to their reference (no reference referring to an area or referring to a basic amount). **Charts** (Figure 104), including quantity symbols (Figure 105), **dot maps** (Figure 106) and **isolines** (Figure 107) are suited to represent absolute data (no reference). However, automatically generated dot maps are not suitable, because the dots are distributed randomly or schematically across the area. Meaningful dot maps therefore require quite elaborate production methods.

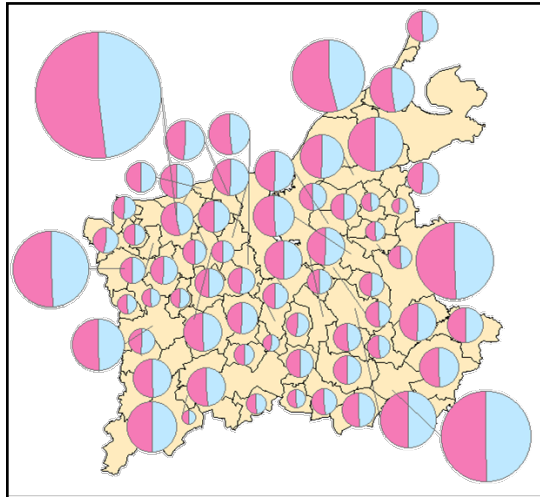


Figure 104: Charts

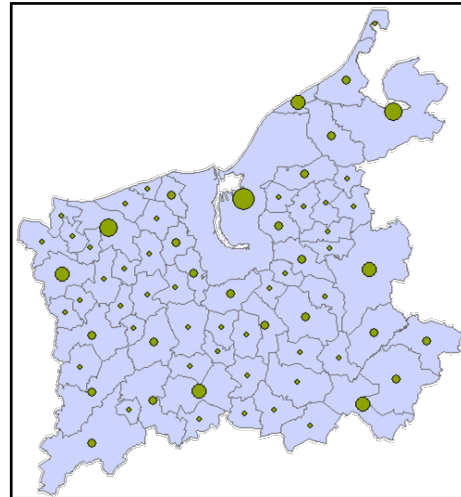


Figure 105: Quantity symbols

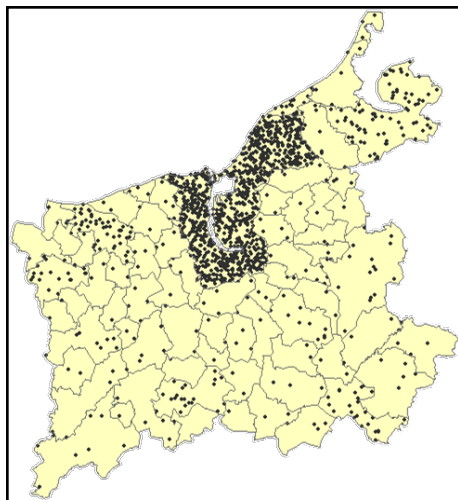


Figure 106: Dot map (automatically generated)

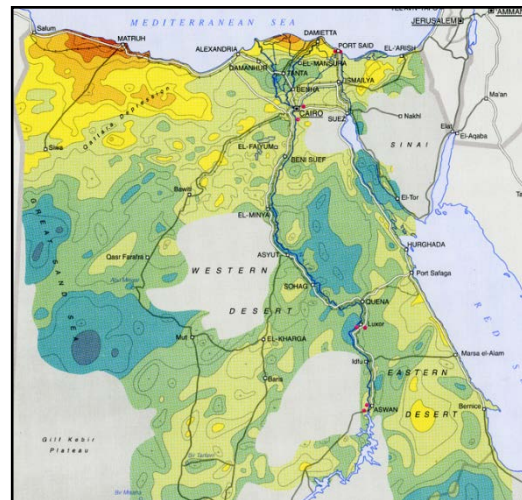


Figure 107: Isoline map

Relative values are represented with **choropleth mapping** (Figure 108). The areas that will be coloured are either the actual reference areas (if the data is referring to an area) or the signature symbolising the basic amount (if the data is referring to a basic amount). The visualisation of data referring to a basic amount in a computer-assisted way is possible, but it requires more manual steps than the classical choropleth map with data referring to an area. The influence of the reference area with data referring to an area becomes obvious when comparing the absolute values, used to calculate the reference values. In the following example two communities (A and B) are compared regarding the total number of inhabitants and the resulting population density.

Community B has more inhabitants than community A, which is larger than community A (Figure 109). Considering the population density another conclusion may be drawn (Figure 110).

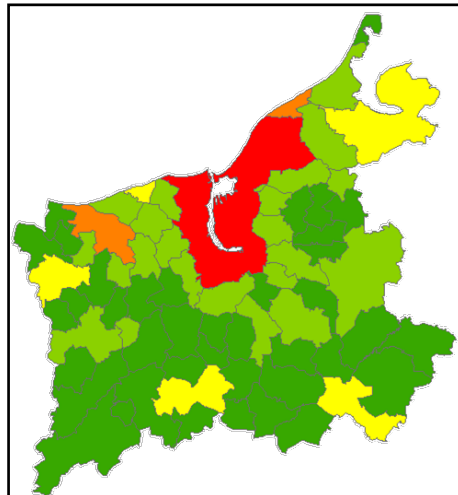


Figure 108: Chloropleth map

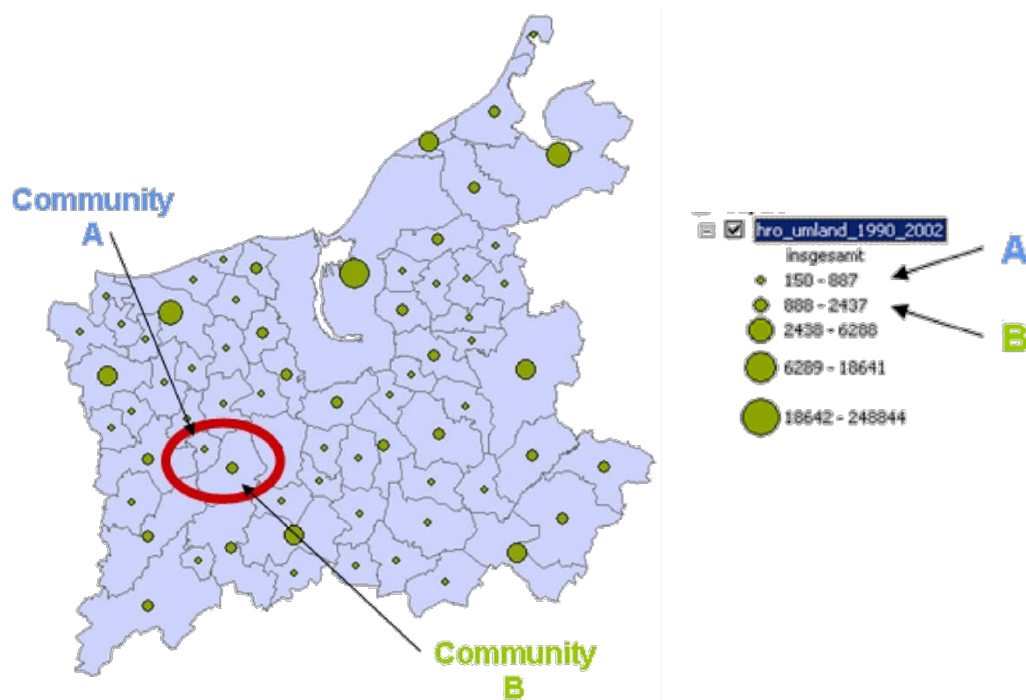


Figure 109: Total number of inhabitants

Despite the smaller total number of inhabitants, community A has a higher population density than community B. The size of reference areas may falsify the map statement. A possible solution to meet this effect is the usage of regular reference areas. All geometric shapes which can be composed to an area-covering graticule are suited to this task (e.g. triangles, rectangles or hexagons). Figure 111 shows the population density of Germany calculated based on  $10 \times 10 \text{ km}^2$  squares.

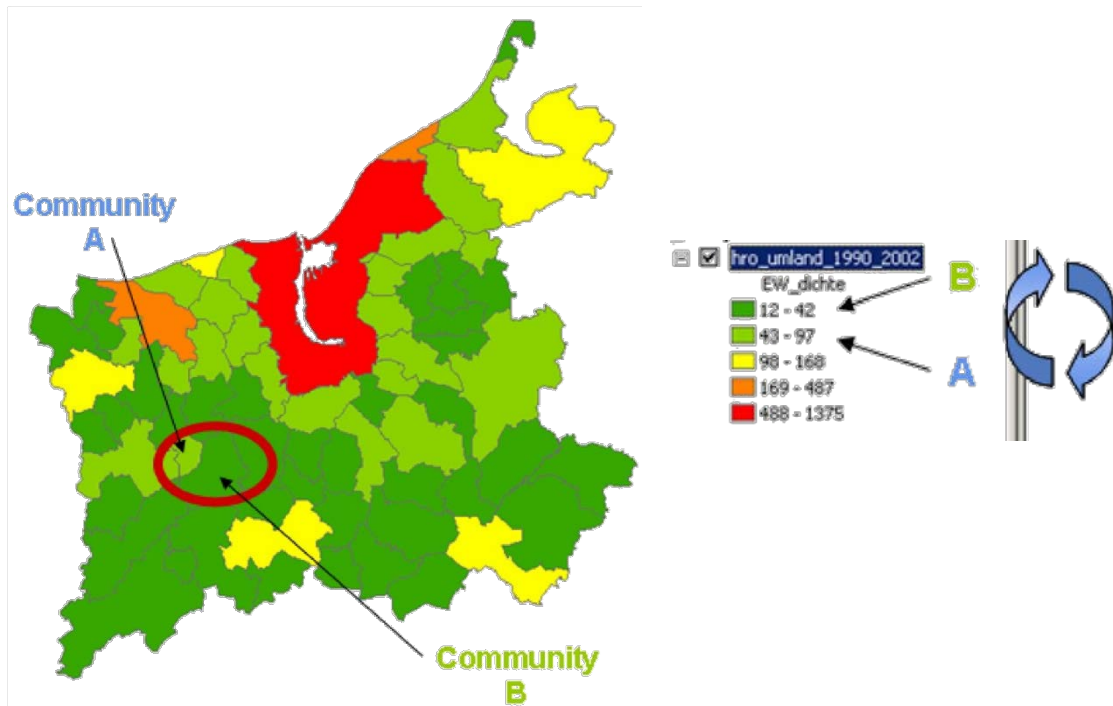


Figure 110: Population density

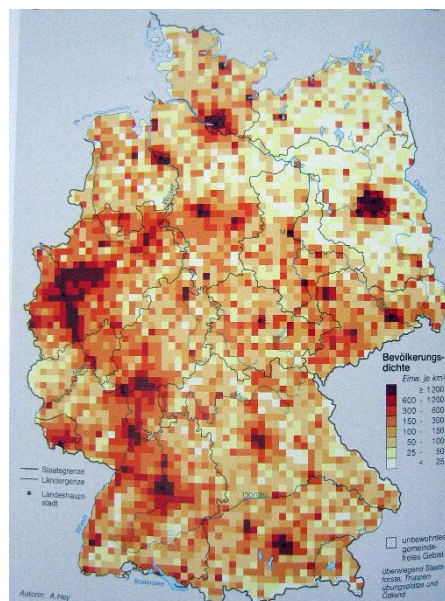


Figure 111: Population density of Germany based on 10x10km² squares

Qualitative data, i.e. data which answers the question “what?”, are visualised with point or linear symbols (Figure 112). Objects that stretch over areas are represented using a qualitative area symbolisation. Maps which are designed with qualitative area symbolisation are called “**mosaic maps**” (Figure 113).



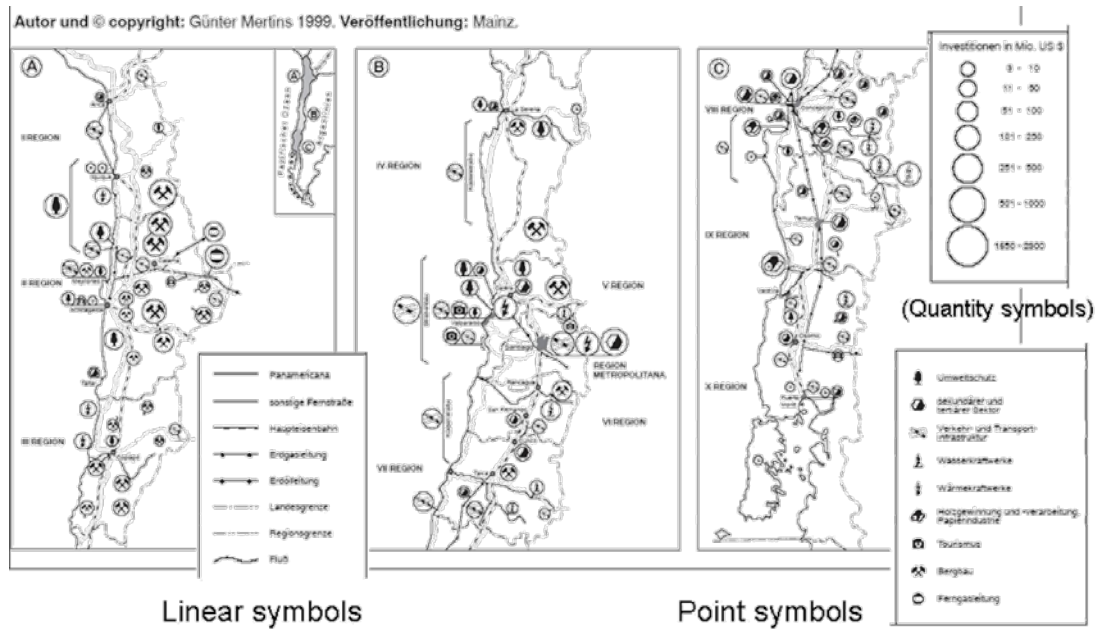


Figure 112: Point and linear symbols (Mertins, 1999)

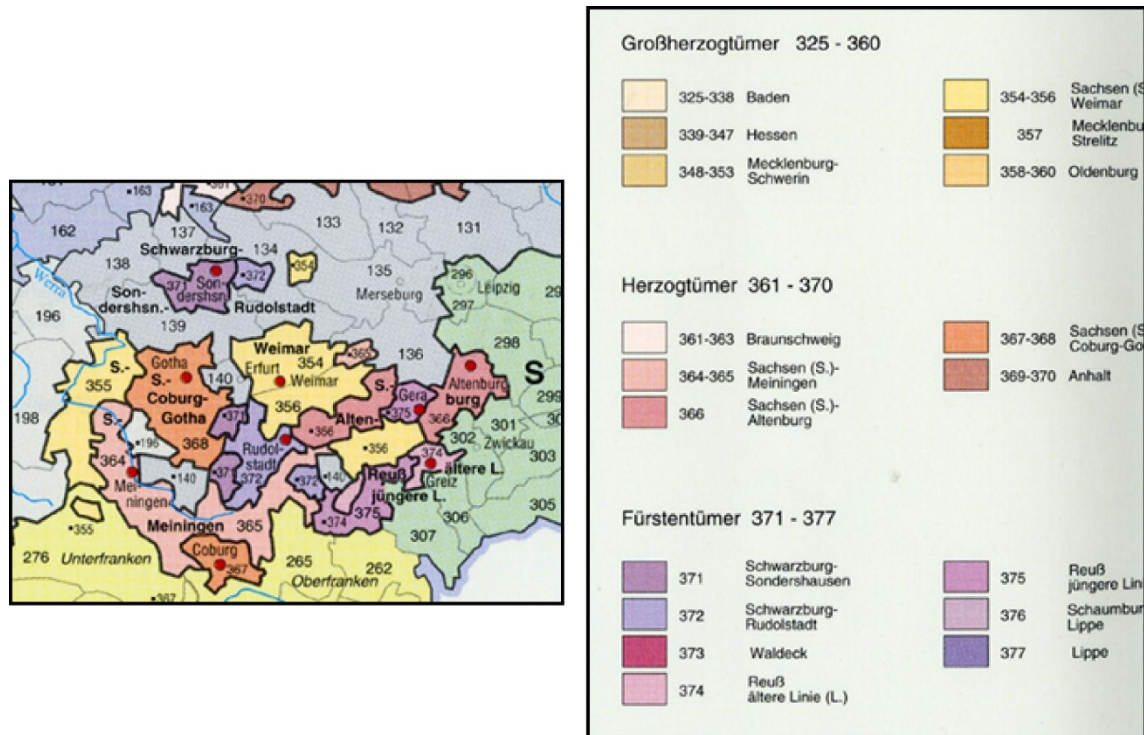


Figure 113: Qualitative area symbolisation (mosaic map)

Directions and movements, whose representation should answer the question „where?“ are visualised using **vectors** ( ).

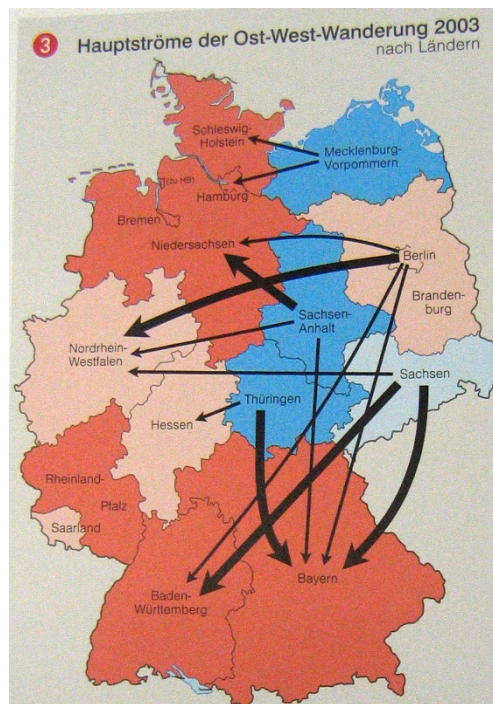


Figure 114: Migration flows represented using vectors

## 7 Map-related representations

### 7.1 Introduction

To create more vivid representations of geographical space, map-related representations are often employed. These should allow even inexperienced map-users to gain information on the depicted area. Map-related representations are particularly often used in the tourism sector.

**Definition:** Two-dimensional, perspective, artistic, and pictographic terrain visualisations, which differ from conventional maps are referred to as ‘map-related representations’.

Examples of map-related representations are shown in Figure 115.

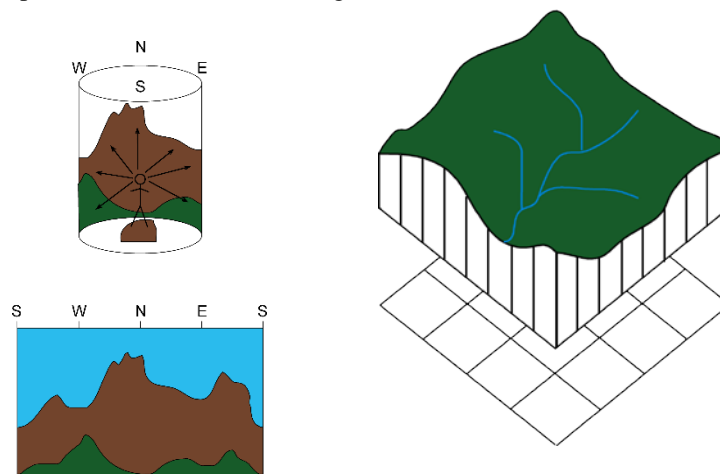


Figure 115: Map-related representations – (a) panorama (central perspective) and (b) block diagram (parallel perspective)

**Reliefs** (Figure 116) and **globes** (Figure 117) are three-dimensional map-related representations.

**Definition:** A scaled physical model of parts of the terrain is called ‘relief’. In scales smaller than 1:50 000 the relief is banked to avoid flattening the terrain and therefore a loss of three-dimensional

significance. Relieves with imprinted or on-painted map information are referred to as ‘map relieves’ (Buchroithner & Stams in Bollmann & Kock, 2002).

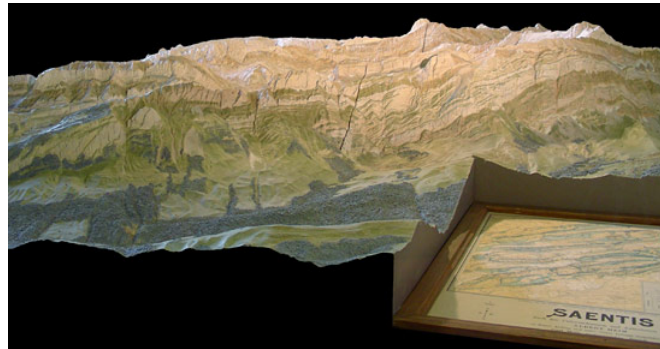


Figure 116: Relief (source: <http://www.terrainmodels.com>)

**Definition:** A globe (lat. ball, sphere) is a three-dimensional map-related representation form of the Earth and other celestial bodies as well as the ostensive celestial globe. In most cases it is a model of the real world, on which the global theme is depicted area-wide and undistorted (Hiller, 2005).



Figure 117: Globe

## 7.2 Special types of topographic maps

Topographic maps may be produced in special forms for particular purposes. In general, these special types of topographic maps are not defined as thematic maps. Special types of topographic maps are topogram, aerial photograph maps and satellite image maps.

**Topograms** are highly schematic representations where only single points are in a correct positional arrangement. A topogram only includes objects which are relevant to its purpose. It is an incomplete representation of geographical space. The most important attribute of topograms is the topologically correct representation. A typical example of topograms is a route network plan of public transport (Figure 118).

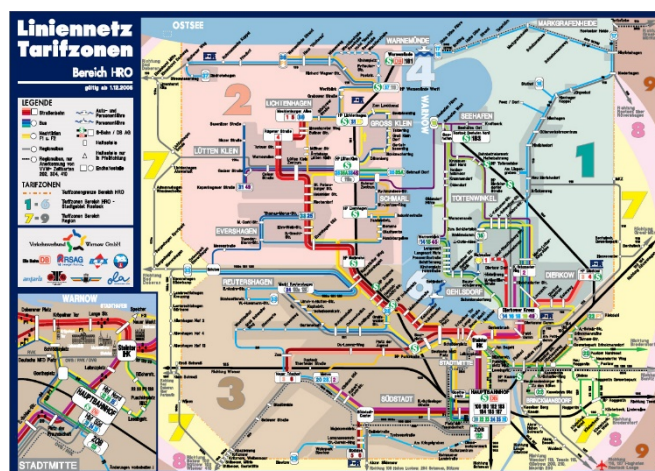


Figure 118: Topogram (route network plan of public transport in Rostock)

**Aerial photograph maps** (Figure 119) and **satellite image maps** (Figure 120) are produced by combining selected map information with an aerial photograph or a satellite image. The image must first be transformed into the map

projection. These image maps often are more vivid than classical topographic maps, but the cartographic information is much reduced.



Figure 119: Aerial photograph map



Figure 120: Satellite image map (source: Google Maps)

## 8 Digital cartography

As in all fields of life, the development of computers has caused changes and opened new possibilities in the field of cartography. From analogue map production, computer-assisted map production and digital cartography have evolved.

### 8.1 Introduction

The purpose of digital cartography is an economically-efficient production, design and revision of significant, cartographically correct and up-to-date maps (topographic and thematic maps) as well as other cartographic products. Digital cartography replaces analogue procedures. The vision is a map that is largely computer-generated. The purpose is the production of several different maps from one data set, adjusted to the demands of the individual map user.

These individual maps should be produced as economically as possible whilst meeting high quality standards. With a computer the attributes, spatial distribution and the alteration of objects can be visualised based on geographical data. Because of the use of methods of computer graphics, the quality of a computer-generated map is *nearly* independent from the knowledge and experience of the map editor. Although the influence of the 'human factor' is disappearing from the map-making process, it cannot be neglected. Not every theoretically possible representation method is useful as well as every possible combination of map themes in thematic maps. The critical eye of the map editor and, if necessary, the corrective intervention, is still essential.

Several types of software are used for digital map production. Besides graphic software (e.g. FreeHand, CorelDraw), CAD-software (e.g. AutoCAD, GeoCAD), GIS-software (e.g. ArcGIS, MapInfo, QGIS) and cartography-software (e.g. CartoDB, MaPublisher) are also used. Graphic software focuses on a pleasing design and offer a great number of graphical design versions. The software has almost unlimited possibilities, which may lead to bad maps if the editor is lacking knowledge and experience. **Graphic software** is well suited to transforming scanned analogue maps into digital vector format. **CAD-software** allows a geometrically exact representation. Originally, GIS-software focused on data capturing and data analysis. Today's **GIS-software** packages possess extensive presentation tools that offer a lot of automatic functions. As seductive as all these functions may be, the editor has to have a critical eye on the product and, if need be, must choose a more suitable representation function. The possibility of manual intervention is decidedly limited in GIS-software compared to graphic software. **Cartography software** tries to combine the advantages of graphic software (free design and manipulation) and GIS-software (storing data in a database; separate geometry from representation). Another advantage of cartography software compared to graphic software is the focus on cartographic design, which allows special ways and techniques of representation (e.g. dotted-dashed line).

The most important advantage of digital map production is the possibility of 'testing' different representation methods without great expenditure of time. Due to the numerous possibilities of presenting the data there is a danger of adapting the map to the possibilities of the software instead of optimising the map layout according to the map's theme. Digital map production increases efficiency in the map making process.

## 8.2 Colour models

Colour models were developed to explicitly describe colours. Names such as 'green', 'purple' or 'aquamarine' are not clear and therefore not suited to explicitly defining colours. Colour models have been developed to solve this problem.

The best-known colour models are the **model of light colours (RGB)** and the **model of surface colours (CMY)**. Both models vary in the method of colour mixture. The model of light colours works with additive colour mixture while the model of surface colours uses subtractive colour mixture. **Additive colour mixture** means that coloured spectra of light are added if overlapping. The overlapping of all coloured spectra of light produces white light. Based on this idea all colours are split into the colour components red, green and blue. The RGB-model is applied to all 'luminous' media. All kinds of screen (e.g. computer monitors) belong to this group.

The model of surface colours works with a **subtractive colour mixture**. Due to an overlapping of the basic colours cyan, magenta and yellow, one range of the spectre of light is absorbed at one time (range of the complementary colour). This process can be illustrated as follows: a blank sheet of paper is white. If cyan is printed on this paper it will appear in a light blue. From the whole spectre of white light the range of red is absorbed by cyan. If the process is continued with yellow printed on the cyan paper the paper will appear green. Now the red and the blue range of the spectre of white light are absorbed. The only remaining visible range of light is green. If another layer of magenta is added, no range of the visible spectre of light will remain. The paper will appear black. The model of surface colours is used in the printing industry. The addition of black (**CMYK**) is made to avoid problems when thin black lines are to be printed. It is very difficult and therefore expensive to print three thin lines in cyan, magenta and yellow absolutely coincidentally. Besides this, the overprint of three times a hundred percent covering colour will make the paper become undulated (due to the wetness). Black is therefore introduced as a special colour, which is printed in a separate step.

Colour models can be visualised by **colour figures**. There are several colour models, which work with **brightness, hue and saturation**. The best way to represent these models is a colour double cone (Figure 121). The colour double will be explained using the example of the **HSB colour model (Hue, Saturation, and Brightness)**.

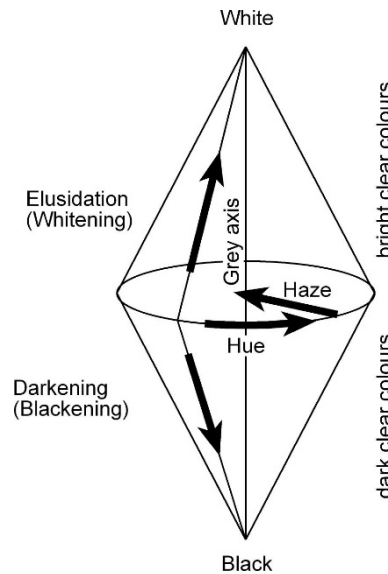


Figure 121: Colour double cone

The circular area where both cones meet represents the hue. It changes clockwise or counter-clockwise. Starting from the middle, the saturation of hue increases towards the edge. The axis connecting the apexes (representing black and white) visualises the brightness. The upper cone contains bright clear colour, while the lower one contains the dark clear colours. Besides the colour double cone there are other colour figures such as the colour cube and the colour sphere. A cut through the colour sphere reveals a well-known colour representation – a colour circle.

To be able to transform colour definition from one model (e.g. colours of a graphic chosen on screen → RGB) into another model (e.g. colour for printing → CMYK), reference models are used. The scope of reference models is to define certain attributes of colours metrically. In 1931, the CIE (Commission Internationale de l’Eclairage = international illumination committee) determined such a reference model, meant for international adoption – the **XYZ-model**. For this model the tristimulus (standard colour value) was determined according to the sensitivity of the 3 colour receptors in the human eye. The three components were each assigned to a colour (X = red, Y = green and Z = blue). Nearly 40 years later, the xy-model was derived from the XYZ-model, forming the basis of the CIE-chromaticity diagram (Figure 122).

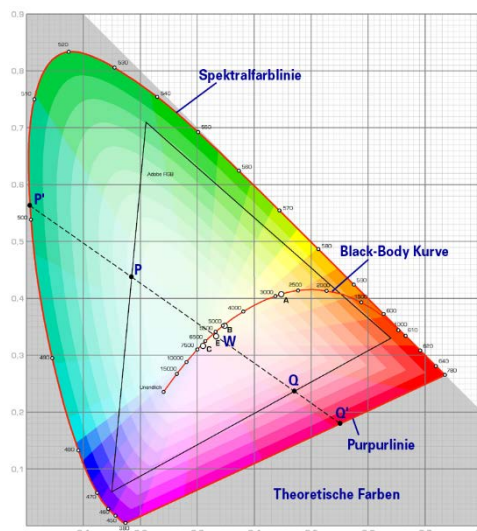


Figure 122: CIE-chromaticity diagram (source: Wikipedia/Torge Anders)

This **CIE-chromaticity diagram** (often referred to as the ‘sole of a shoe’) should visualise all colours that can be perceived by the human eye. The range of colours that can be presented on a screen is limited to a triangle inside the coloured area. This triangle is unique for every screen. This range of colours presentable by a device is called the ‘**colour space**’.

The **xy-model** was enhanced forming the **Lab-model**. This model bases on the segmentation of a colour into a brightness component (L) and two colour components (a,b). The colour component a contains the green and red portions of a colour. The colour component b contains the blue and yellow portions. Today, the Lab-model is

widely spread and often employed. The brightness component is also called luminance whilst the colour component is also called chrominance.

When choosing or evaluating colours, it should always be taken note of that colour reproduction always depends on the device. This can be done by regularly calibrating screens and using colour profiles for scanners and printers or plotters. Colour profiles contain information on the individual colour space of the device and a transformation rule for the usage of reference models. Some software allows the definition of colours by giving colour coordinates in RGB or CMYK. For the planned output on a printing device it is recommended to define the colours in CMYK right from the beginning. Tables of colours may help to choose the right colour in CMYK when working on a device with a RGB colour space. Tables of colours systematically summarise different shadings of colour components (CMYK) or hues (Pantone). To some extent, the colours are printed on different kinds of paper to illustrate the alteration of colour effects depending on the print media.

If the graphic (e.g. the map) should be displayed on screen then the colours should be defined in RGB. For colours in the WWW a list of web colours has been created that are defined using a hexadecimal code. An overview of this list can be found in the internet (e.g. <http://www.mywebsite.force9.co.uk/web-colors/color-guide.htm>) The colour space of the user's screen is out of reach for the map editor and therefore remains an uncertainty factor.

### 8.3 Raster and vector data in digital mapping

Digital graphic data has either raster or vector format. The main difference between those formats is in the structure of data. Vector data is stored as points, lines and areas (like the geometry data in a GIS). Raster data is organised in picture elements (pixel = raster cell). To each pixel, a certain data value is assigned. Due to the structure of the data, raster and vector data have different attributes.

**Vector data** needs less memory than raster data because only some point coordinates and relations between them have to be stored. The resolution of a vector data set is not limited which means that zooming can be done without a loss of quality. The easy scaling allows different output sizes. The vector format is suited to store graphic data that consists of points, lines and even coloured areas. In particular, digital maps belong to this type of data.

In contrast, **raster data** has a fixed resolution determined by the number of pixels. To some extent, zooming without a visible loss of quality is possible. When raster data is enlarged, the existing pixels are stretched to the new format. To avoid gaps, the pixels are simply multiplied. Because of this, the pixel structure becomes visible when zooming in too much. The storage of data values for each pixel of raster data needs more memory. There are lossy and loss-less methods of compression. The raster data format is suited to store graphic data that contain colour gradients (half-tone images). Particularly photos (aerial photographs, satellite images) belong to this type of data. The background of aerial photograph maps and satellite image maps is an image (raster data). The map geometry is a layer in vector format.

Editing raster data, also called '**image processing**', is the manipulation of objects that consist of pixels. Methods used for image processing are among others filter operations and the manipulation of image channels (mostly colour channels in RGB). These methods are primarily employed in remote sensing. In the field of cartography, the manipulation of raster data is limited to manual cosmetic corrections. The advantage of raster data is the editing and analysis basing on pixels. Disadvantages are the fixed resolution of the image that is determined by the pixel-size and the difficulties in creating and editing objects caused by the pixel structure. Objects can be created by using pattern recognition. The following example illustrates the effect of the fixed resolution. In the first zoom level (Figure 123) the outlines are sharp. The pixel structure of the image is invisible.

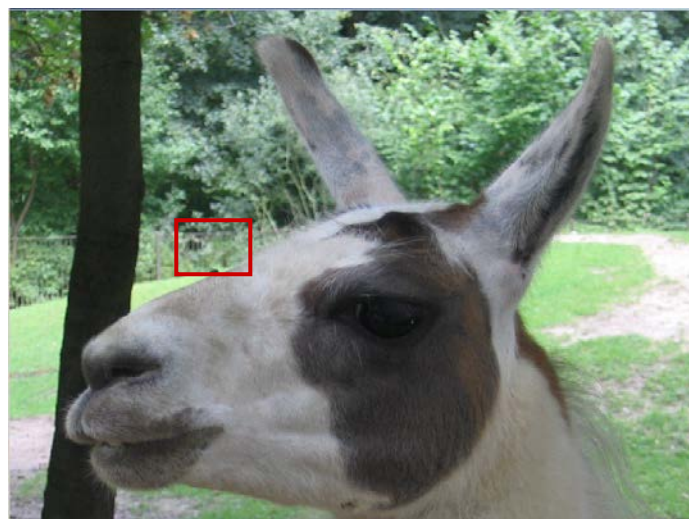


Figure 123: Raster data; normal zoom level (source: A. Hey)

At a high zoom level, the pixel structure of the image becomes visible (Figure 124). The block structure arises from the multiplying of pixels that is required to avoid gaps in the image.



Figure 124: Raster data; high zoom level (source: A. Hey)

The editing of vector data (computer graphics) bases on another approach. The geometry of objects is mathematically described by curves and/or straight lines. This allows the editing of single objects. The resolution of the graphic is not fixed. The data can therefore be presented in different sizes without or with just a few changes. Due to the logical delimitation of single objects it is possible to link attribute data to them. In this way, for example, polygons (areas) representing districts can be linked to numbers of inhabitants. To use this advantage of vector data a GIS-software is necessary – it cannot be done in normal graphics software.

Figure 125 and Figure 126 illustrate the flexible resolution of vector data: even intense zooming-in does not change the clear outlines.

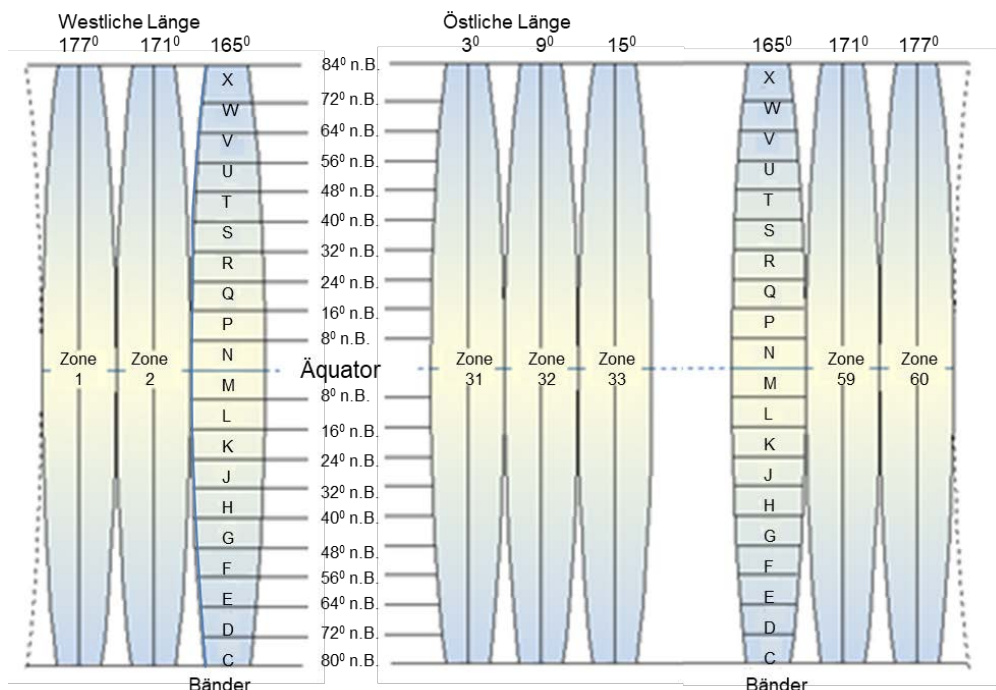


Figure 125: Vector data; low zoom level (source: A. Hey)



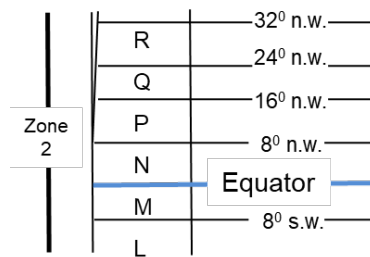


Figure 126: Vector data; high zoom level (source: A. Hey)

## 8.4 Raster data formats in digital mapping

There are several raster data formats, each with different specifications. Table 10 summarises some of the best-known raster data formats and lists their most important attributes. A main attribute of raster data formats is the compression. Lossless methods allow a reduction of file-size without a loss of data, which means a compressed image can be restored without a loss of quality. With a lossy method of compression (e.g. JPEG-Format) higher compression rates can be achieved, but the loss of quality may be considerable.

Due to its internal structure, the **TIFF-format** (Tag Image File Format) depends on the platform it is created on. This means it cannot be transferred from PC to Apple Macintosh or vice-versa. The reason for this is the different sequence in which the bits are stored (so-called big-endian or little-endian). TIFF-images are organised either in stripes or in tiles. This speeds up the image set-up (short load times). It is possible to calculate image pyramids (pre-calculated zoom levels) which further increases the speed. The TIFF-format supports colour depths up to 24 bit, which corresponds to approximately 16.7 million colours. This format is therefore well-suited for half-tone images such as photos. Because the TIFF-format stores uncompressed photos, the files are very large, which is why the TIFF-format is not generally suitable for use in the internet.

Format	File-suffix	Quality loss caused by compression	Colour depth	Maximum number of colours
TIFF (Photo)	*.tif	without	1,4,6,8,12,24 Bit	16.7 million
GIF (Map)	*.gif	without	1,4,8 Bit	256 (colour-palette)
PNG (Map, Graphic)	*.png	without	16 to 48 Bit	16.7 million
JPEG (Photo)	*.jpg	Minimal to considerable, according to compression level	24 Bit	16.7 million

Table 10: Raster data formats

The **GIF-format** (Graphics Interchange Format) limits the maximum image size to 16,000×16,000 pixels. The resolution is not stored, so the number of pixels determines the image size according to the resolution of the screen. A maximum of 256 colours can be stored. These colours are not predefined (indexed colours). For each image a unique colour table is created. If an image possesses more than 256 colours (e.g. a photo), colour gradients are simulated by a method called 'dithering' (Figure 127).



Figure 127: 'Dithering' in a GIF image (source: Wikipedia)

The method of dithering uses adjacent colours for adjacent pixels to create colours that are in-between. Because the human eye cannot disintegrate colours geometrically exactly it 'internally' creates a half-tone. Due to this colour mixing, the contrast in the image is reduced. GIF is especially often used with animations. The GIF-format allows several images to be stored in one file, which then can be used like a flip-book to create an animation. Examples of GIF-animations can be found online<sup>29</sup>. Because of the lossless compression method and the reduced colour depth of just 8 bit, GIF-files are very small in size and therefore well-suited for use in the internet. The GIF-format should be employed for figures with colour gradients, such as graphics with clearly defined homogeneously coloured areas (e.g. digital maps).

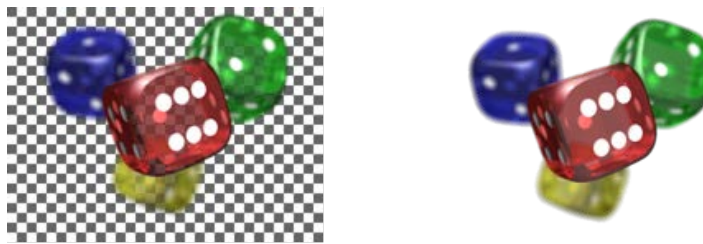


Figure 128: PNG-images using transparency and masking

The **PNG-Format** (Portable Network Graphics) differs from the GIF-format in the supported colour depth. The PNG-format supports up to 4 channels. Besides the three colour channels (RGB), an additional alpha-channel is available, which can be used to store masks (Figure 128). Masks are necessary if a figure should have a shape other than a rectangle. Due to the pixel matrix, every raster image is rectangular. The PNG-format can store 256 transparency levels. It is platform-independent and uses a lossless compression algorithm. As for the GIF-format PNG, does not store a resolution. The disadvantage of PNG compared to GIF is that PNG allows only one image per file. PNG can therefore not be used to store animations. Because of the relatively small file size, the lossless compression and the supported colour depth, the PNG-format is suited for all kinds of images in the internet.

The **JPEG-format** (Joint Photographic Group) is especially well-suited for images with colour gradients (half-tone images, e.g. photos). It is not suited for line drawings. The reason for this is the implemented compression algorithm (wavelets) with a flexible compression rate. The method is very effective, but with an increasing compression rate, block artefacts appear. Even colour patches are disturbed by this as they appear 'spotted'. The JPEG-format is therefore not suitable for vector graphics (do not mix up with vector data!), like digital maps. The main application for of the JPEG-format is for photos. Because of the small file size, JPEG images are very well-suited for the internet. The following example illustrates the effects of compression of JPEG images. The original image (Figure 129) has a size of 2.92MB. The highly compressed image (Figure 130a) is reduced to 236kB. A closer look on an image detail in the compressed image reveals the block artefacts (Figure 130b).

<sup>29</sup> e.g. [http://de.wikipedia.org/wiki/Bild:Rotating\\_earth\\_%28small%29.gif](http://de.wikipedia.org/wiki/Bild:Rotating_earth_%28small%29.gif)

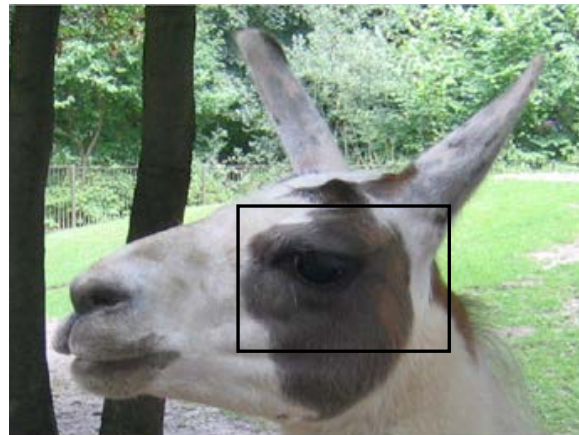


Figure 129: JPEG-image; original (source: A. Hey)

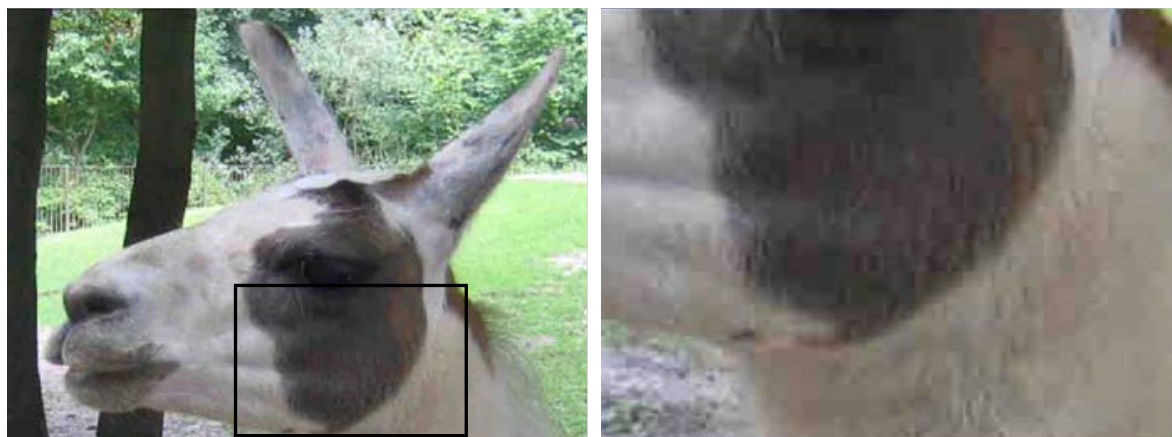


Figure 130: JPEG-image; (a) highly compressed and (b) showing block artefacts (source: A. Hey)

## 8.5 Digital data capture

The aim of digital data capture is the creation of vector data which can be modified and adapted to different output formats very easily. To use an analogue original as the basis for producing a digital map, the analogue map first has to be transformed into digital data. Digital data may still have to be transformed into another digital format so that it can be used as a digital map. There are several methods that can be applied for this:

- scanning (analogue original → raster data)
- digitising (analogue original → vector data)
- vectorising (raster data → vector data)

Analogue maps that are not too large can be scanned easily. During the **scanning** process, the map is optically scanned and transformed into digital information. According to the type of the original map, the scan parameters have to be defined. Black and white originals (so-called line originals) should be scanned with a higher resolution than coloured originals. The colour depth has a great influence on the file size of the scanned image. It should not therefore be set too high. Consider a line original scanned with a colour depth of 24 bit: the file would be more than 20 times as large as necessary. There would also be colour pixels in the scanned image producing a noisy image.

The colour depth bases on the basic sizes of the internal memory structure of a computer. 1 bit stores the information 0 or 1. If several bits are combined, more values can be stored. A cluster of 8 bits (1 byte) corresponds to 256 values (from 0 to 255 =  $2^8$ ). These are 256 brightness-levels when applied to the colour depth (0 = black, 255 = white). More than these 256 brightness levels can hardly be perceived by the human eye. To store colours, 3x8 bits are combined (colour depth = 24). According to the RGB colour model, 256 intensity levels of each colour (red, green, blue) are stored. This gives a total number of 16.7 million colours. The required colour depths according to the type of scan original are:

- |  |                              |
|--|------------------------------|
| • line original (black/white):                       | 1 Bit                        |
| • continuous tone original (greyscale picture):      | 8 Bit                        |
| • colour original without colour gradients (figure): | 8 Bit (with indexed colours) |

- colour original with colour gradients (photo): 24 Bit

The term ‘**digitising**’ describes the manual transformation of an analogue original into digital format. The digitising can either be done using a digitising tablet or directly on screen (desktop mapping). **Manual digitising** on a digitising tablet offers the advantage of being able to deal with large-size originals, assuming an accordingly large digitising tablet is available. Single objects can be selected for digitisation. In this way, a generalisation can be performed while digitising. The disadvantages of using a digitising tablet are the to some extent high costs of the required digitising tablet and the lack of zoom facility. A digitiser tablet consists of a desktop, where a wire graticule is integrated, and a digitiser mouse with magnifying glass. The wire graticule records the movements of the mouse and captures the local coordinates of the digitised points. The denser the wire graticule, the higher the resolution of the digitiser tablet. The magnifying glass of the mouse is equipped with crosshairs and serves for the input of points. It is not possible to zoom-in more than the magnifying glass allows. Besides the high costs of a digitising tablet, the space requirement should not be neglected.

The advantage of **on-screen digitising** compared to the above mentioned way of digitising is primarily the clearly lower need for disk space of the computer. Similarly to digitising on a digitising tablet, the on-screen digitising offers the opportunity to perform a generalisation simultaneously with the data capturing. The workflow of desktop mapping starts with a scanned analogue map that is used as digitising background. The digital vector data is then captured with the mouse in one or more layers on top of the raster map. The digitised data can immediately be checked regarding completeness and correctness. When working with a digitising tablet then a control print is necessary to detect such errors. On screen ‘blind areas’ may be magnified by zooming in. The only limit is the resolution of the background raster data. The disadvantage of desktop mapping is the limitation of the visible digitising area on screen to a relatively small part of the digitising area due to the screen size.

For a **vectorisation**, a raster image already exists. The transformation is performed automatically or semi-automatically with special software. Vectorisation needs less manual steps than digitising on screen. Special software allows the conversion of raster into vector data. However, the original data has to meet certain requirements regarding contrast and complexity. Simple lines, such as isolines or water lines, can be captured quite easily and rapidly. The software allows choosing the level of accuracy of the data capture. The disadvantages of vectorisation are the high effort, which are necessary with complex originals, and the rather limited possibility of performing a generalisation during the data capturing.

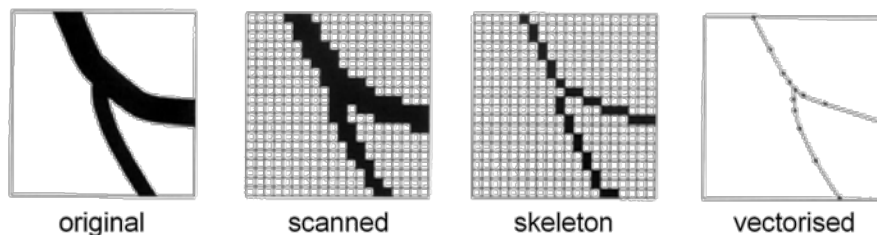


Figure 131: Steps of vectorisation (source: Olbrich et al., 2002)

An example of simple lines is used to illustrate the method of vectorisation. After the scanning the resulting raster image is skeletally thinned until the objects (lines) are not thicker than 1 pixel. Thus the connection from one pixel to another is explicit. Often the software needs to be told by an operator where to start with following one line. Starting from this given point the software records the local pixel coordinates and finally forms a line out of the pixel chain (Figure 131).

**Automatic digitising** allows a fast conversion of raster into vector data with flexible accuracy levels. Algorithms for automatic vectorisation dispose base on pattern-recognition methods. Complex maps may require an immense effort in post-processing.

For each individual case it should be considered which method of digital data capture is best suited. Desktop mapping can be performed without extra equipment, just a computer and graphics software. The other methods all require additional equipment.

Digitising may produce errors that will lead to problems in further data manipulation and data analysis. All captured data has therefore to be checked for errors. Typical errors that arise from digitising are:

- lines are interrupted by gaps → undershoots
- lines intersect → overshoots
- several lines do not meet in one node → node cluster
- due to a deviation from the straight line, lines appear to be tiered

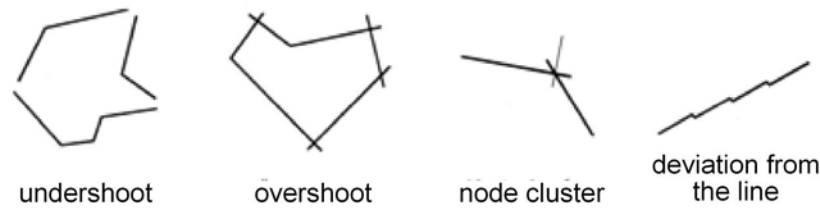


Figure 132: Errors arising from digitising (source: Olbrich et al., 2002)

When planning the production of a map including existing data, it should be noted that the conversion from vector into raster data is always possible in an automated way while the other direction, from raster into vector data may be much more difficult. This is why enough time should be planned to perform the raster vector conversion manually.

Digital maps can be created using GIS-software. This offers the opportunity of linking geometry with attribute data. The presentation on the basis of attribute values may then be partly automated. Geometry data are available in vector format. These vectors are assigned attributes according to an object catalogue. For example, key numbers of land use are assigned this way. By this step the geometry is separated from the representation – in the data capturing and data analysis all vectors look the same.

The graphic format is taken from a **signature catalogue** and automatically assigned to the appropriate vector. The representation is therefore very flexible and may be changed within a moment by using another signature catalogue. Due to the linked attribute data, thematic maps can be created quite automatically by selecting the objects to be visualised in the map according to their attribute values. For example a road map may be produced this way, by only selecting vectors with the attribute ‘road’.

**Check: There is only one object class catalogue, but any number of signature catalogues!**

## 8.6 Multimedia maps

The structure of digital screen-maps differs from ‘classic’ analogue maps regarding the fact that a screen map can include different types of media. According to Bill (1996), media depending on time and media independent from time can be distinguished (Figure 133). Especially the integration of time-dependent (dynamic) elements is impossible in analogue maps. Analogue maps can only contain static media as elements of the map frame or on the back of the map. **Time-dependent media** comprises animations, videos and audio-files. Media considered to be **independent from time** (static), are texts, tables, graphics, photos and 3D-representations. These different types of media are not essential for a map, but they can reasonably complete it.

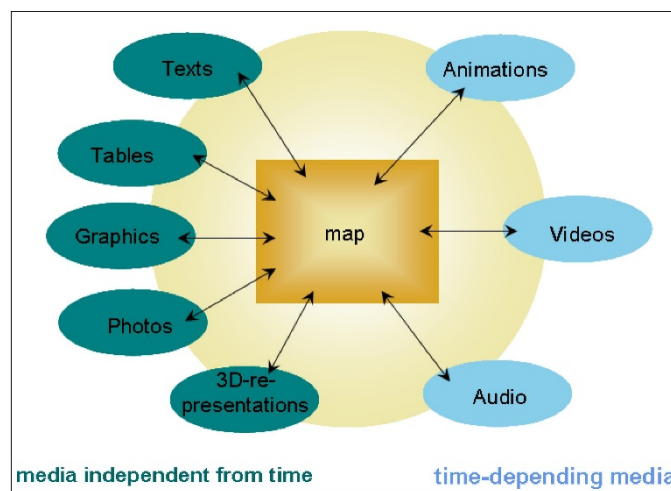


Figure 133: Elements of screen maps (Bill, 1996)

Types of multimedia maps are:

- animated maps
- internet maps
- digital 3D-visualisations (e.g. virtual reality)

According to Buziek et al. (2000), animations can be divided into temporal and non-temporal animations. **Temporal animations** are visualisations that change in the course of time, such as the traffic levels during a day.

**Non-temporal animations** are visualisations of movements in space, such as fly-throughs. Animations can be calculated in real-time or completely recorded beforehand.

**Internet maps** vary in the level of interactivity they allow. According to the audience these digital maps possess different attributes. The types of internet maps according to Olbrich et al. (2002) are:

- static maps
- dynamic maps
- interactive maps

**Static maps** cannot be altered. Mostly they are raster maps which are included in the web page as simple images in image formats such as png or jpeg. **Dynamic maps** are employed to visualise developments. Similar to a film, trends are presented using an animated temporal series. A typical example of dynamic maps is the 'satellite film' in the weather forecast. **Interactive maps** offer a lot more possibilities for the user to interact with the map. In simple versions of interactive maps, additional information can be reached by clicking in the map. These maps are called '**sensitive maps**'. An example of a sensitive map is GoogleEarth. Here, photos of certain places may be shown by clicking on points in the map.

Complex interactive maps allow the selection of the visible thematic layers in the map. Also the way of representation or the map detail, which shall be represented, can be determined by the map-user. These individually created maps are referred to as '**maps on demand**'. Many consider them to be the map type of the future.

## 8.7 3D visualisations

3D visualisations work on the principle of **stereo vision**. Because of the distance between the eyes, a person sees two slightly different images. The images differ in the viewing angle. Basing on this small difference, the brain calculates distances to the perceived objects and creates a spatial effect. The principle of stereo vision uses this attribute of human vision. An observer gets a visualisation where several pictures taken from slightly different viewing angles are included in a way that one eye will always see just one picture. The two images perceived at one time form a stereo-pair. This method can be realised digitally as well as analogue.

An analogue 3D-stereo-method that has been in use for a long time is the **anaglyph method**. In this process, a stereo pair (two images of the same objects taken from slightly different positions) is printed one above the other in complementary colours (normally blue-green and red). To gain a 3D impression from this anaglyph, special anaglyph glasses are needed. The lenses of these anaglyph glasses are in complementary colours. Only one of the images is visible to each eye at a time. The principle of glasses which show only one image per eye at a time is also used with **digital stereo representations**. For this, special glasses are synchronised with a screen. The glasses can be separately blinded by polarisation of the in-built liquid crystals. The screen shows alternating images, while the glasses are blinded accordingly. Because the change happens very fast, the user does not perceive the shutting of the glasses.

Another analogue method is the **lenticular foil technique** (Figure 134). Using this method, true 3D visualisations can be created. True 3D means, that no aids are necessary to gain the three-dimensional impression. For the lenticular foil technique, images from different positions (different viewing angles) are taken. These images are cut into small stripes and combined through a complex procedure into one image. On this image a foil consisting of terete lenses (also called lenticular lenses), is applied. The lenses cause refraction so that only one image per eye is visible. The difficulty of this method is the exact positioning of the foil according to the stripes of the basis images. If more than one stripe is visible at a time or if the visible stripes do not belong to the same image the resulting image will be blurred.

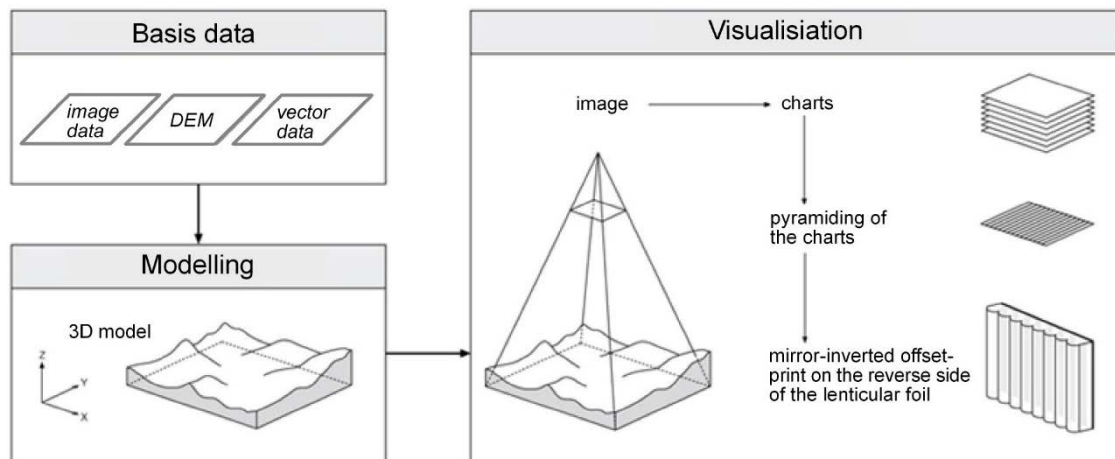


Figure 134: Lenticular foil technique (from www.mbmsystems.de)

With the lenticular foil technique, new application fields are open to cartography. The representation of relief becomes very vivid, but also the production of multi-lingual maps is possible, by changing the visible labels with the changes of the viewing angle. Even different images can be combined, so that changes in the course of time may be visualised (to a limited extent). The representation of map details by zooming-in is available as well. These map details are only visible from certain viewing angles, so that the whole map will be visible as well.

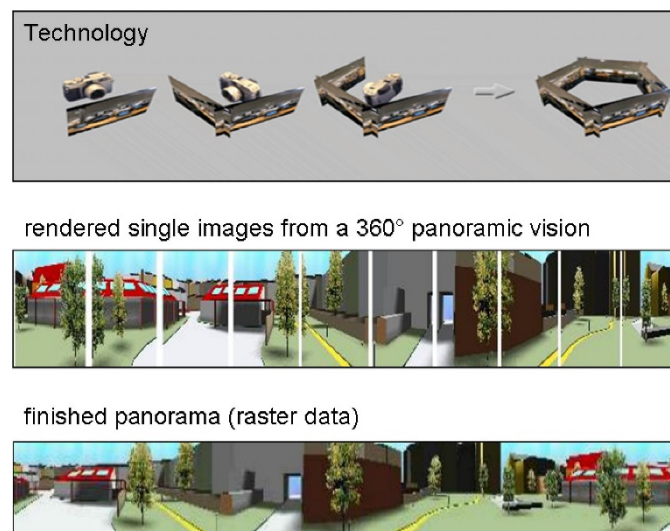


Figure 135: Creating a panorama

**Virtual reality** is a type of digital 3D-visualisation. The term ‘virtual reality’ stands for a technique that offers people the possibility of communicating with a computer and of visualising and manipulating complex data. The term was introduced in 1989 by Jaron Lanier. It contains an oxymoron: while ‘virtual’ means something physically not existing, ‘reality’ represents something verifiable and revisable. Technologies employed in virtual reality applications are for example the virtual reality modelling language (VRML) for the 3D vector world and Quicktime Virtual Reality for the 3D raster world. The virtual reality approach may be combined with a classic GIS. For this, the data storage and the functions of manipulation, selecting and analysing data are similar to those of a normal GIS. The user-interface and the opportunities to interact with the computer are designed using virtual reality techniques. One way to create a virtual reality environment is to create a **panorama** and from this producing a 3D-impression (Figure 135). First, several images from one position are taken with the camera being rotated through a small angle for each photo. In the image processing the overlapping areas are then used to connect the images. Basing on these linked images, the panorama (raster data) is calculated.

The virtual reality technique may be extended to create an **augmented reality**. This is a new kind of human-visual-interaction where information is superimposed into the user’s field of vision, e.g. by using data glasses. The superimposing is contextual, which means that the information is derived from and fits to the observed object, e.g. a building. The real field of vision of a potential buyer of real estates may thus be enhanced with information on free estates in the observed area or building.

Another method of digital 3D-visualisation are **3D city models** (Figure 136). The model can cover a whole city or it can be limited to a quarter or even single buildings. In 3D city models, attribute data is linked to corresponding geometry data similar to GIS. 3D city models may be employed for interactive walk-throughs or for analysis (e.g. change detection). There are different detail levels (Level of detail 0, 1, 2, 3, and 4) according to the depicted area and the purpose of the model. Simple models show buildings as simple blocks. These so-called ‘block models’ (LOD 1) use only a little information, such as ground plans, building heights and streets as basis. Roof types are not considered. This simple version of a 3D city model is applied to scales from 1:10,000 to approximately 1:25,000. Block models place lower demands on hard-/software than the extended block model. The extended block model needs more detailed information on buildings. Additionally, ridge lines (roof types), simple front textures and vegetation are included (LOD 2). This method is much more elaborate and therefore it is used only in large scales (approximately from 1:500 to 1:10,000) and for small areas, such as individual city quarters.

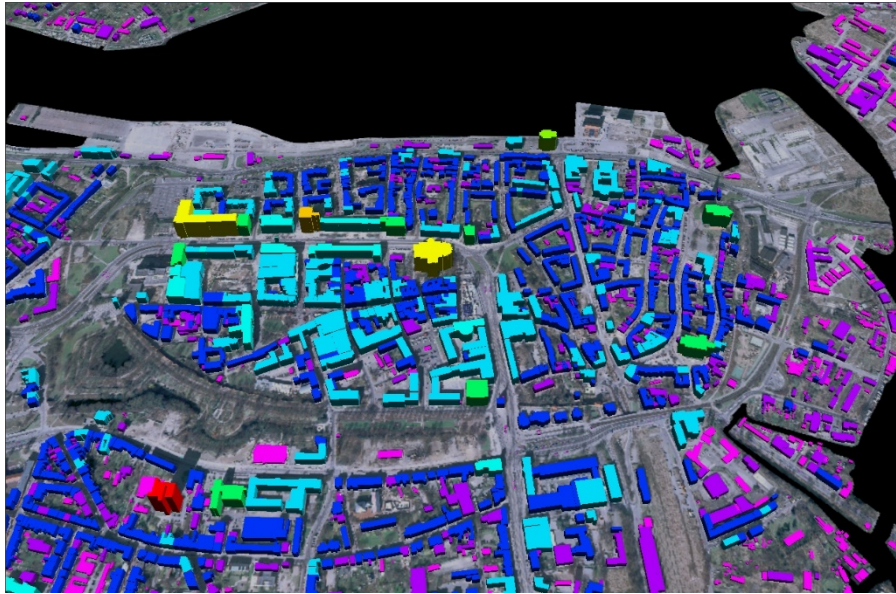


Figure 136: 3D-city model of Rostock (source: Chair of Geodesy and Geoinformatics, University of Rostock)

For all types of 3D city models the DTM (Digital Terrain Model) may be employed. The information needed to create a 3D city model can be obtained by photogrammetry or laser scanning. Cadastral and planning data may also be used. 3D city models are employed in many fields, such as city planning, mobile radio, providers, environment, tourism, assurances, real estates and architecture. 3D city models may appear more realistic when applying textures (Figure 137).



Figure 137: 3D-model including textures and vegetation (source: diploma thesis Kai Bannicke, 2004)

Visualisation of geographical data will in future develop from today’s level (few interactions, knowns being published in maps) to a high degree of interactivity and the exploration of unknown relations between the data and individual selections.

## References

Bertin, J. (1983): *Graphische Semiologie. Diagramme, Netze, Karten.* Walter de Gruyter, Berlin.



- Bollmann, J., Koch, W.-G. [editors] (2001): Lexikon der Kartographie und Geomatik, Band 1. Spektrum Akademischer Verlag, Heidelberg.
- Bollmann, J., Koch, W.-G. [editors] (2002): Lexikon der Kartographie und Geomatik, Band 2. Spektrum Akademischer Verlag, Heidelberg.
- Buziek, G., Dransch, D., Rase, W.-D. [editors] (2000): Dynamische Visualisierung, Grundlagen und Anwendungsbeispiele für kartographische Animationen. Springer Verlag, Berlin.
- Hake, G., Grünreich, D., Meng, L. (2002): Kartographie. Walter De Gruyter, Berlin.
- Olbrich, G., Quick, M., Schweikart, J. (2002): Desktop Mapping, Grundlagen und Praxis in Kartographie und GIS. Springer Verlag, Berlin.
- Töpfer, F. (1979): Kartographische Generalisierung. Ergänzungsheft Nr. 276 zu Perermanns Geographische Mitteilungen, VEB Hermann Haack, Geographisch Kartographische Anstalt Gotha/Leipzig.
- Parry, R. B., Perkins, C. R. (2002): World Mapping Today. Butterworth-Heinemann, Oxford, UK.
- The American Heritage Dictionary of the English Language (2007). Houghton Mifflin, Boston, MA, USA.

## **Part E**

# **Information Systems and Databases**

PD Dr. Meike Klettke, M. Sc. Markus Berger and Prof. Dr.-Ing. Ralf Bill



## 1 Introduction

Information systems store and administer large amounts of data. Geographical information systems normally store coordinates (geometry/topology) and additional data concerning the real world. Such data is usually stored in files. If several applications need to share the same information then a software component developed specifically for storing and querying that data can offer great advantages. A **database system** is such a software component. These systems store different kinds of data in a structured, efficient way and offer a standardized language (SQL, structured query language) to access it.

Database systems are made for storing large amounts of data. They are able to query millions of data sets very efficiently. For this, they use index structures and different strategies to optimize access to the data with minimal response times. Database systems can also realise data security (protect data against loss) and data protection (protect the data against unauthorised access or changes).

In this chapter, the fundamentals of **relational databases** are introduced, representing their data as table structures. Furtheron SQL (**Structured Query Language**), the database query language, is also summarised. This chapter also shows how to apply these kinds of databases in geographical information systems.

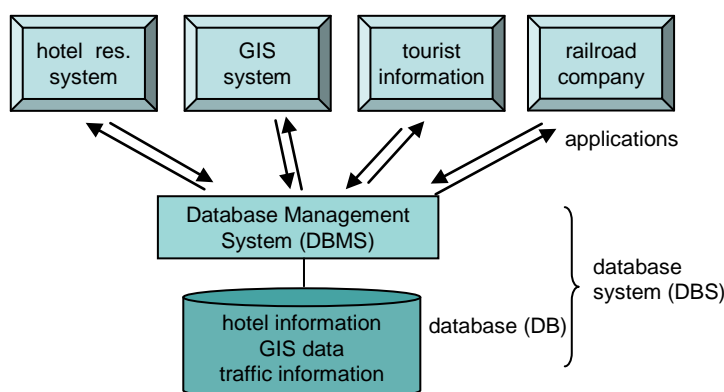


Figure 138: Typical architecture of an information system

Figure 138 shows a typical architecture for a complex software system that contains a database component. The database system is generally an internal component: the user does not get to see it, only other applications communicate with it.

**Definition:** A **database system (DBS)** consists of a **database management system (DBMS)**, a software product for realising all functionalities of the database system and, additionally, one or more **databases (DB)** that contain the data belonging to a database system. There are different interfaces for communicating with a database system: the **Structured Query Language (SQL)** supports creating the tables (CREATE), filling a database with data sets (INSERT), updating (UPDATE) and deleting (DELETE) data sets and querying (SELECT) the stored data. SQL is standardised and all available database systems support this language.

Several commercial (e.g. ORACLE, Microsoft SQL Server) and free/open (e.g. PostgreSQL, MariaDB) database management systems exist that can be used in software products. The concrete database design and model has to be developed separately for each application. For this reason, the focus in this chapter is on designing new database schemata, inserting and updating data sets and querying the information in the database. Additionally, some functionality that is available in all database management systems is introduced briefly.

## 2 The relational data model

The relational data model was suggested by Edgar Codd (1970). Even today many, if not most, database systems use this model. It is widespread and can be applied for different kinds of applications. In this section, the foundations of the relational model are introduced.

### 2.1 Structure

A relational database contains **relations** (also called *tables*) that contain **attributes** (or *columns*). The data sets in the relations are the so-called **tuples** or *rows* of the table. Figure 139 shows a sample relation and illustrates the basic terms of the relational data model.

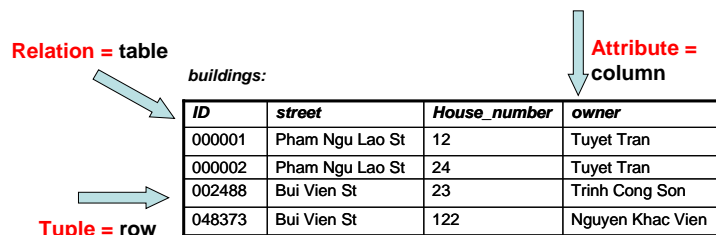


Figure 139: The relational data model

Each attribute has a **data type**. The data type describes which kinds of values can be stored in that attribute/column. Available data types in database systems are, for example, integer numbers, decimal numbers, dates, times or enumerations. More than 40 basic data types are available in all databases systems.

Figure 140 shows two sample relations with some data that are also in relation to each other. Different data types (string and integer values) occur in these relations.

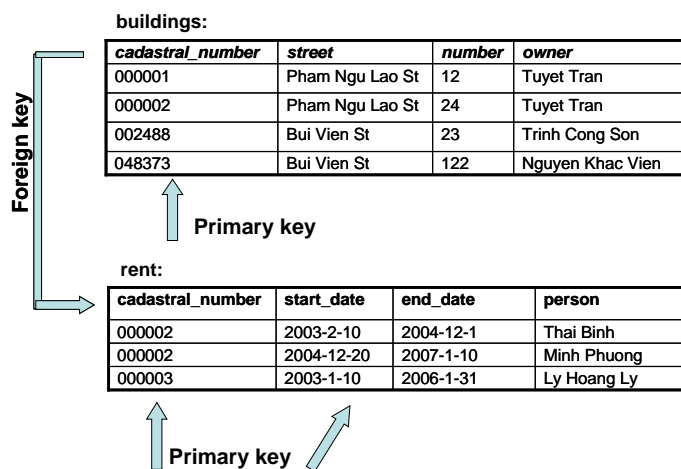


Figure 140: Sample relations with integrity constraints

## 2.2 Integrity constraints

The attribute or the set of attributes that identify each row in a relation is called the **primary key** of the relation. Typical examples for primary keys are unique numbers, for instance persons identification number, students enrolment number or other IDs. Also, names can be unique (for instance product names). In other cases, a combination of attributes can uniquely identify a row. For example, the combination of a student number and the university name is unique, and thus we can store a reference to each student with just these two attributes. If an attribute or a set of attributes is a primary key of a relation, then there may not exist two tuples with the same values for the primary key attributes. In Figure 140, the attribute `cadastral_number` is the primary key of the relation `buildings`. The relation `rent` has two attributes that together identify the tuples: the `cadastral_number` and the `start_date` because only one person can rent a building at a particular time.

Relations can contain so-called **foreign keys**. Foreign keys are attributes or sets of attributes that only contain values that occur in another relation and that are primary keys of the other relation. In Figure 140, the attribute `cadastral_number` is a foreign key of the relation `rent` and references the primary key of the relation `building`.

## 2.3 Operations

The relational model defines tables as the only data structure for storing information. Additionally, operations on the tables are defined. These operations are used for querying the data and an individual result table is generated from the original data structure. The basic operations on relations are selection, projection, Cartesian product, join, union, and rename.

The operation **Selection** chooses data sets of a relation that fulfil a special condition. **Projection** (unrelated to map projections) reduces the number of columns in the result table. The **Cartesian product** combines all data sets of two or more different relations. The **Join** operation combines the information stored in different tables over attribute characteristics, for instance over attributes that have identical values. **Union** collects the data sets of two relations with identical number of attributes and compatible data types in one relation. The operation **Rename** is

easy to understand: it changes an attribute name. With all these operations the resulting database status is generated corresponding to a query of the user. Details about the query language can be found in section 5 of this chapter.

### 3 Conceptual design

Databases for real applications tend to become quite large. They can contain hundreds of different tables. For designing larger databases, a **conceptual model** is used which supports the user in defining the main objects, their characteristics, and their relationships, but hides technical details. The conceptual model is applicable for communicating with domain experts and discussing general interrelations of the domain. This information is the prerequisite for a correct database design.

#### 3.1 Application of a conceptual model in the database design process

Figure 141 shows the integration of a **conceptual model**, in this case the **entity-relationship model (ERM)**, in the design process.

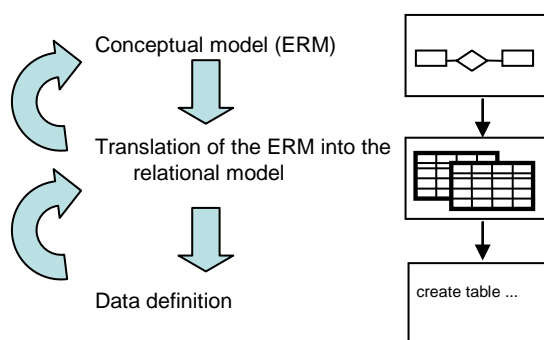


Figure 141: Database design process

The design process starts with a conceptual model. This model is translated into a **relational model** that can be seen on the second level. The next step is the creation of the tables in the relational database system, for which a special language is available that will be introduced in section 3.3. But first, the conceptual model is described and introduced with an example.

#### 3.2 The entity-relationship model

A very popular and widespread conceptual model for database design is the entity-relationship model suggested by Peter Chen (Chen (1975)). Figure 142 shows a small example for the design with this model.

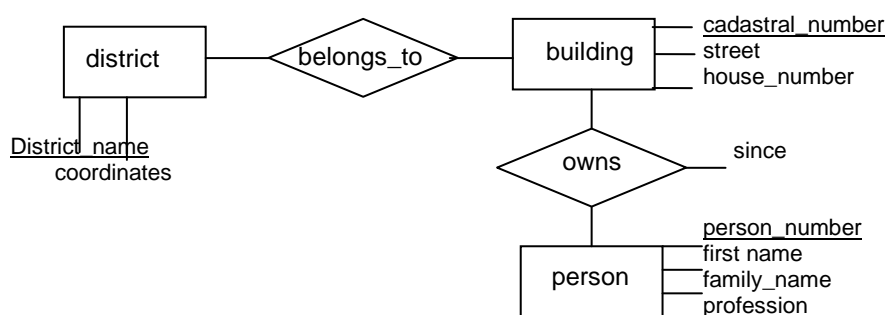


Figure 142: Sample entity-relationship model

In this model the basic **entity types** (here district, building and person) of the application are identified and defined. The **relationship types** between the entities (belongs to, owns) are also defined in the model. These types model the database on the table-level. Individual entries (tuples) in a table would be called **entities** and **relationships**, but are not part of the conceptual model.

The sample model shows how to model buildings that are associated with geographic information. The entities in the sample database are buildings, persons that own a building, and buildings belonging to a district. The relationship between the buildings and persons describe which person owns which building, while the relation between buildings and districts describes which building belongs to which district.

Entities and relationships can be described by attributes that contain characteristics of the entities or of the relationship. The primary keys of entities have to be defined. Primary key attributes are those attributes or attribute

sets that can uniquely identify an entity. For this example, the entity type `person` has an attribute `person_number` that is unique for each person and can be used for identification. This primary key attribute is underlined in the entity-relationship model. The `building` type has an associated `cadastral_number`: we assume that such a unique number exists in the cadastral register. Each `district` has a name: there are no different districts having the same name, which is why the name of the district can be used as the primary key of this entity type.

Another piece of information that is necessary in the conceptual model are **cardinalities** which specify the strength of relationships between objects. A cardinality constraint is defined by two values: the minimum and the maximum value. The **minimum value** determines the minimum number of times that an entity participates in a relationship. Typical values are 0 (relationship is optional) and 1 (relationship is mandatory). The **maximum value** determines the maximum number of times that an entity participates in a relationship. Typical values are 1 (at most one relationship) or n (n indicates any number equal to or greater than the minimum).

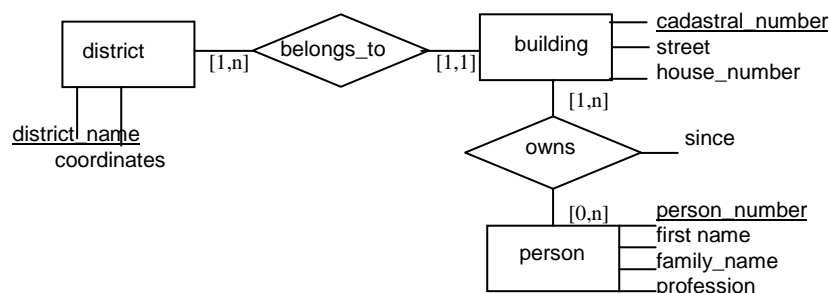


Figure 143: Sample entity-relationship model with cardinalities

Figure 143 shows the sample entity-relationship model with cardinalities. The cardinality `[0,n]` between `person` and `owns` means that while a person can own one or more buildings, it is also possible that a person does not own a building. More formally, a person owns at minimum 0, at maximum n (that means any number) buildings. The cardinality `[1,n]` between `building` and `owns` means that a building is owned by at least one person, it can also be owned by several persons (maximum n).

The cardinality `[1,1]` between `building` and `belongs_to` expresses the fact that a building belongs to exactly one district (minimum 1, maximum 1). The last cardinality in the example is the cardinality between `district` and `belongs_to`. This cardinality is specified with `[1,n]` which means while at least one building belongs to a district, the maximum number is n - typically many buildings belong to a district.

### 3.3 Translation into the relational model

The entity-relationship model with entity types, relationship types, attributes, cardinality constraints and primary key information can be translated into the relational model: the corresponding tables are generated in this process. Entity types are translated into relations, and all attributes of each entity type are translated into attributes of its relation. Data types are not part of the entity-relationship model and need to be set during the translation process.. The primary key of the entity type becomes primary key of the relation. In Figure 144 the relations (tables) that are generated in the translation process are shown.

person	<u>person_number</u>	first_name	family_name	profession
building	<u>cadastral_number</u>	street	number	
district	<u>name</u>	coordinates		

Figure 144: Translation of the entities into relations

The translation of the relationships is a bit more complicated. Most relationship types translate into relations that contain the attributes of the relationship type, as well as the primary keys of the associated entity types. For example, if we translate the relationship type `owns` into a relation then this relation has three attribute: the attribute `since` (of the relationship), the attribute `person_number` that is the primary key of the entity `person` and the attribute `cadastral_number` that is the primary key of the entity `building`.

The attributes `person_number` and `cadastral_number` are foreign keys: `person_number` references the primary key in the relation `person`, and the attribute `cadastral_number` references the primary key in the relation `building`.

The last question that arises is: which attributes are the primary key of the relation `owns`? It depends on the cardinalities. There are three different cases, which are illustrated in Figure 145.

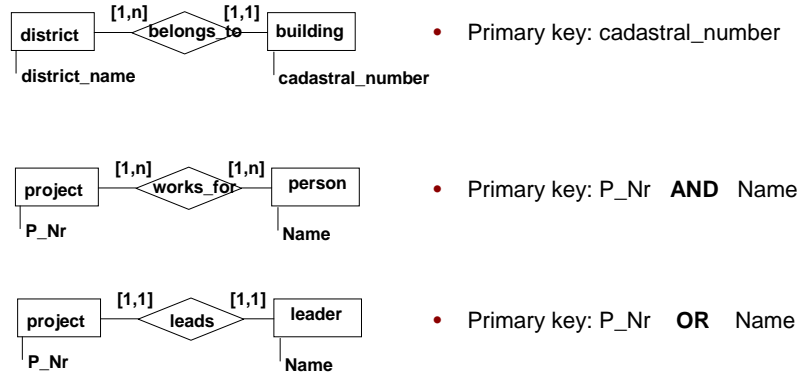


Figure 145: Primary keys for relationships

When using these cases as-is, the two relationship types from the sample entity-relationship model would translate into two relations (as shown in Figure 146).

belongs\_to:

name	<u>cadastral_number</u>

owns:

<u>person_number</u>	<u>cadastral_number</u>	since

Figure 146: Translation of the relationships into relations

However, if the entity-relationship model contains a cardinality [1,1] then the relations generated from the entity types and from the relationship types can be merged. In the current example, the relations for the entity type `building` (with the attributes `cadastral_number`, `street` and `house_number`) and for the relationship type `belongs_to` (with the attribute `name` and `cadastral_number`) can be merged. Both relationships have the same primary key that is a prerequisite for the combination of relations. The result is a new relation that replaces the two relations and contains all attributes of both relations. Figure 147 shows this new relation.

building:

<u>cadastral_number</u>	street	number	district_name

Figure 147: Result of the translation process

This corresponds with case one of Figure 145 and leads to one less relation than in the naïve translation. Case three wasn't part of the sample model but would result in two less relations, as the related entity types are functionally one and the same entity - they require each other, and can thus be merged.

The **Data Definition Language (DDL)** is used to implement these relations in a database system. With this language it is possible to define which table to generate, which attributes to put into the table, their types, and its primary and foreign keys. The following DDL statements generate all relations from Figure 145, Figure 146 and Figure 147.

```
create table district (
    district_name    varchar(10) not null primary key,
    coordinates      blob)

create table building (
    cadastral_number integer not null primary key,
    district_name    varchar(10),
    street            varchar(30),
    house_number     integer,
```



```
foreign key (district_name) references district(district_name))
```

```
create table person (
  person_number integer not null primary key,
  family_name   varchar(30) not null,
  first_name    varchar(30) not null,
  profession    varchar(40))
```

```
create table owns (
  cadastral_number integer not null,
  person_number    integer not null,
  since            date,
  primary key (person_number, cadastral_number),
  foreign key (person_number) references person(person_number),
  foreign key (cadastral_number)
    references building(cadastral_number))
```

In these statements we can see that the primary key and the foreign keys are also defined. All information in the DDL statements is derived from the entity-relationship model.

### 3.4 Normalisation

Relations that contain redundant information can cause several anomalies. Figure 148 provides an example for this.

<i>cadastral_number</i>	<i>owner_id</i>	<i>family_name</i>	<i>first_name</i>	<i>phone_number</i>
..	..	..	..	..
23004	1	Tran	Tuyet	032938274
24850	1	Tran	Tuyet	032938274
..	..	..	..	..

Figure 148: Translation of the relationships into relations

If we update one value in this sample, for instance we change the *phone\_number* of the owner in the tuple with *cadastral\_number* '24850' then the relation is not consistent because there are different phone numbers available for one owner. Other persons or other programs accessing the data do not know which information is correct.

<i>owner_id</i>	<i>family_name</i>	<i>first_name</i>	<i>phone_number</i>
..	..	..	..
1	Tran	Tuyet	032938274
..	..	..	..

Figure 149 shows a better solution avoiding redundant storage by splitting the single relation into two relations.

<i>cadastral_number</i>	<i>owner_id</i>
..	..
23004	1
24850	1
..	..

<i>owner_id</i>	<i>family_name</i>	<i>first_name</i>	<i>phone_number</i>
..	..	..	..
1	Tran	Tuyet	032938274
..	..	..	..

Figure 149: Normalised relations

Database theory has defined **normal forms** (1st - 3rd normal form) to avoid such problems and methods for achieving these normal forms. More detailed information can be found in Batini et al. (1991), Elmasri and Navathe (2006) and in Connolly and Begg (2004).

## 4 Inserts of values

The databases created in the database design process contain empty tables. For filling them with data, the following statement can be used:

```
insert into person values (47362, 'Tuyet' , 'Tran' , 'architect')
```

or

```
insert into person (family_name, first_name, profession, person_number)  
values ('Tuyet' , 'Tran' , 'architect', 47362)
```

The first (short) form can be used if all attribute values are in the same order as in the `create table` statement. In the second variant the attributes of the relation are enumerated in the `insert` statement and the values must correspond to the order of attributes in the `insert` statement.

## 5 SQL: Structured Query Language

Once the tables are filled with data, there is a further part of the database language that handles user requests on these tables: the **structured query language (SQL)**. All queries formulated in this language correspond to operations of the relational model, for instance projection, selection, join and so on. This chapter introduces the query part of the language, the basic building blocks of the language and important functions with several examples.

### 5.1 First example for an SQL query

A simple SQL statement consists of the clauses:

```
select ... from ... where ...
```

This chapter starts with an example:

```
select family_name, first_name  
from person  
where profession='architect'
```

As a result this query delivers a table consisting of two columns, `family_name` and `first_name`:

<i>family_name</i>	<i>first_name</i>
Tran	Tuyet

This contains the data sets of all persons that are architects. The query first performs a selection on the rows of the `person` table (selecting all tuples with a `profession` attribute equal to 'architect') and then projects the result by only showing the table attributes `family_name` and `first_name`.

This query shows the general structure of SQL statements. The `from` clause defines which database tables are considered in the query, the `where` clause defines the conditions that the tuples of the result have to fulfil, and the `select` clause defines which attributes are in the result table.

### 5.2 SQL in detail

With an SQL query, a user describes only the structure of the result and not the way in which the result is generated. This characteristic of the query language is called **declarative**.

The `from` clause enumerates which relations are used for generating the result. If in the **from clause** only one table name occurs then the result is generated from this table. It is also possible to enumerate two or more table names. In this case, the **Cartesian product** (all combinations of data sets) of all tuples is generated.

In the **where clause**, the conditions which the tuples of the result have to fulfil are enumerated. This means that the content of this clause acts as a sort of filter over the tuples from the `from` clause. It is possible to compare and thus filter values (of the tables) with constant expressions (for instance: `family_name='Thong'` or `person_number<2000`). It is also possible to compare attribute values with attribute values of other tables, for instance `owns.person_number=person.person_number`.

The **select clause** defines which attributes are shown in the result: every tuple left after the where clause is reduced down to the selected number of columns. The symbol \* can also be used, which means that all attributes from the original table(s) shall occur in the result table.

A more complicated example is the following one. We are looking for the family name and the first name of the owner that owns the house with the address 'Bui Vien St 23':

```
select family_name, first_name
  from person, owns, building
  where (street='Bui Vien St') and (house_number=23) and
        (person.person_number=owns.person_number) and
        (owns.cadastral_number=building.cadastral_number)
```

Result:

<i>family_name</i>	<i>first_name</i>
Cong Son	Trinh

Three relations are used for generating the result: the tables *person*, *building* and the connection between both in the table *owns*. The query constructs the Cartesian product of all tuples, then filters out all buildings that do not have the given *street* and *house\_number*. The *and* expressions demand that all conditions are met (as opposed to *or*, where only one condition has to be met). The third and fourth conditions ensure that only those tuples are selected where the correct person is attached to the building according to the *owns* relation. This is necessary because the Cartesian product on its own includes *all* the possible combinations of these tables, not just the ones that conform to the relationship.

Additionally in the *select* clause and in the *where* clause, so-called *aggregate functions* can be used. The available **aggregate functions** are *sum*, *count*, *max*, *min* and *avg*. *sum* delivers the sum of attributes, *count* the number of data sets, *max* the maximal value, *min* the minimal value and *avg* delivers the average (mean) values.

```
The following query shows an example:select count(*) as number_of_persons
  from person
```

The result of this query is the number of tuples in the relation *person*:

<i>number_of_persons</i>
4

### 5.3 Ordering and grouping of results

Further clauses are available in SQL. The clause *order by* determines the order of the data sets in the result. The following example query illustrates this::

```
select ...
  from ...
  where ...
  order by family_name asc, first_name asc
```

With this example, we determine that the data sets of the result table are ordered by the family names in ascending (alphabetical) order. If the same values for the family name occur, then the data sets are ordered by the first names, also in alphabetic order, as described by the second part of the *order by* clause.

Further possibilities are the *group by* clause and the *having* clause. *group by* summarises data sets with the same values, for instance:

```
select profession, count(*) as number
  from person
  group by profession
```

This query delivers the following result:

<i>profession</i>	<i>number</i>
artist	2
architect	1

computer scientist	1
--------------------	---

Special conditions for groups can be expressed with the `having` clause. The conditions have to be valid for the groups. The following SQL example shows this clause:

```
select profession, count(*) as number
  from person
  group by profession
  having count(*)>1
```

Result:

<u>profession</u>	<u>number</u>
artist	2

The `select` clause specifies the projection list, which means the attributes that are enumerated in the `select` clause occur in the result.

## 5.4 Joins in SQL

The preceding section showed the Cartesian product and how it is used in SQL queries. Another way of combining relations is the use of joins, which implicitly or explicitly put a condition on what tuples are passed on to the `where` clause, instead of starting with all possible combinations. In SQL, join operations are expressed in the `from` clause. For this, the relations that are joined are enumerated together with the condition that has to be fulfilled. The following (a little bit more complicated) example shows a join over three relations. The relations `building` and `own` are joined over the condition `building.cadastral_number=own.cadastral_number`. The result table is the base for the next join operation in the query.

```
select street, house_number, family_name
  from
    ((building join owns on building.cadastral_number=own.cadastral_number)
    join person on own.person_number=person.person_number)
  where (street='Bui Vien St') and (number=23)
```

## 6 Updating and deleting values

Values in the databases can be updated and it is also possible to delete data sets (tuples). The syntax for both operations is shown here with an example.

```
delete
  from building
  where cadastral_number=3
```

The `delete` clause contains a `from` and a `where` clause, all tuples from the relation that fulfil the condition in the `where` clause are deleted. One `delete` statement can delete one or several tuples at once. For deleting all tuples the following statement is used:

```
delete *
  from building
```

The `update` clause on the other hand can change one or several attribute values of a tuple or of several tuples.

```
update own
  set person_number=3, since='2007-09-30'
  where cadastral_number= 1
```

This query changes the owner of a building: the `person_number` of the owner as well as the attribute `since` of the object is updated.

## 7 Characteristics of a database management system

A database management system is a software component that is able to store the information and to generate the results for queries. A database management system internally optimises the data storage and is able to generate results over very large data sets efficiently. Furthermore, a database management system contains components that prevent data loss. Some of the most important tasks of a database management system are:

- Query Optimisation
- Data Security
- Data Protection
- Database Recovery

One of the simplest ways in which a database can optimize queries is indexing. In indexing, there is an operator attached to every attribute in the database tables. That operator defines how values for an attribute are sorted and as such is dependent on the data type of the attribute. For example, the data type of the attribute `person_number` might be 'integer'. In that case, the operator simply sorts values according to their numerical value. If the number instead had a string data type (text) and contained letters as well as numbers, there would be multiple ways to define an operator. Sorting could, for example, be alphabetical and start at the leftmost character of the string, or the individual characters could be assigned arbitrary values so that a sum of each word could be calculated. Regardless of the operator, once we have a definite list of our values, the database system can optimization routines to make the data rapidly accessible and searchable. One example of such a method is to subdivide the sorted list and turn it into a tree structure. More on this and other optimization techniques can be found in textbooks such as Batini et al. (1991), Elmasri and, Navathe (2006) and Connolly and Begg (2004).

## 8 Types of database systems

As mentioned at the start of this chapter, relational database systems are an invention from the 1980ies. And while they are still used to manage many of the databases in the world, other types of databases have become increasingly important over the years. These alternatives usually developed for specific use cases that relational systems cannot cover efficiently, but some have even started to overtake relational system for large, general-purpose databases deployed on a global scale.

There are too many variations to discuss in detail here, but we will still try to provide an overview. One of these non-conventional systems, and likely the most important for the readers of this book, are the so-called **spatial database systems**. They are usually extensions to relational database systems, solving their troubles with geospatial information. Much of this trouble derives from the realities of indexing – conventional relational databases assume one-dimensional values, in that a collection of values can be sorted one after the other in an unambiguous way. This is not the case for spatial references, which are generally two- or three-dimensional in nature. For example, the database system could sort a two-dimensional geographic coordinate point either by its latitude or by its longitude – both cases would not make much sense in terms of query optimization. If we wanted to query which points are close to a point of our choosing, only comparing them on one axis is not sufficient, because they might be close on one axis, but on the other side of the globe in the other. If on the other hand we were to abandon indexing all together, we would have to compare every point in our database with every other point for each spatial operation that we perform. Examples of spatial operations are buffers, reprojections, and area calculations.

The way spatial databases solve this is by adding **two- and three-dimensional indexing strategies**. Where a conventional database would organize a list of values into a simple tree structure, spatial indexes subdivide the whole coordinate space (the surface of the earth) into hierarchical, rectangular parcels of different sizes. These parcels grow smaller and more detailed the more features gather in one place. They are still two-dimensional, but because they are strictly hierarchical, the database system can sort them into a tree structure just like one-dimensional values. If we then want to do a spatial query, we can very efficiently reduce the number of points we need to query by first identifying which parcels are even affected (which is easy because operations on rectangles in a two-dimensional space are trivial), and then only doing comparisons over the very reduced subset of points that are in those parcels.

Another problem for relational databases are the highly dynamic and unstructured forms of data often found in multimedia applications. Image, videos, or even live-streams are difficult to store in the strict confines of an environment consisting entirely of table structures. Usually, relational database systems can only store references to these contents. And not only multimedia data is becoming more unstructured – in times of Big data, data mining and user-generated content, most of the generated data does not necessarily fit neatly into a table with strictly typed attributes. Database systems that are able to handle these kinds of inputs are as numerous and varied as the data itself, but are generally referred to as **Not Only SQL (NoSQL)** databases (Harrison, 2015). Examples of these systems include:

- **Document-oriented database:** Information is stored in semi-structured documents, such as XML or JSON. Each object has attribute values, but nature and number of these attributes is usually not enforced.
- **Key-Value database:** Based on a data structure usually known as **dictionary**. Every object is referenced by a key and one or multiple values are attached to each key. These values can consist of everything from numbers to references to other keys.

- **Graph database:** A database following concepts from graph theory, where data and data relations translate to edges, nodes, and properties. This is especially efficient if the focus of the database is on data relations instead of the data itself.
- **Stream databases/Stream data management systems:** This database system is able to manage continuous data streams, such as sensor data, social media messages etc. The query for data is continuously executed until it is explicitly uninstalled. Since most stream databases are data-driven, a continuous query produces new results as long as new data arrive at the system.
- **In-memory databases** primarily relies on main memory for computer data storage in contrast with other DBMS that employ a disk storage mechanism. Thus, they are faster than disk-optimized databases because disk access is slower than memory access, the internal optimization algorithms are simpler and execute fewer CPU instructions. Accessing data in memory eliminates seek time when querying the data, which provides faster and more predictable performance than disk

Each of these systems has several variants, and there are also systems that fall even outside of these three examples. Increasingly, NoSQL databases are deployed in cloud infrastructures, where they deal with input from a multitude of applications and distributed data storage in data centres scattered all over the world.

Check: However, even in the light of all this, relational database systems still have their advantages: What they lack in flexibility, they make up in structure. SQL permits users to construct powerful queries with simple language, while working with NoSQL databases usually require the user to have at least rudimentary programming experience.

## References

- Batini, C., Ceri, S., Navathe, S.B. (1991): *Conceptual Database Design: An Entity-Relationship Approach*, Addison Wesley, 1991.
- Chen, P. (1975): *The Entity-Relationship Model: Toward a Unified View of Data*, Proceedings of the International Conference on Very Large Data Bases, September 22-24, 1975, Framingham, Massachusetts, USA.
- Codd, E.F. (1970): *A Relational Model of Data for Large Shared Data Banks*, Communications of the ACM, volume 13, number 6, pages 377-387.
- Connolly, T.M., Begg, C.E. (2004): *Data Base Systems: A Practical Approach to Design, Implementation and Management (4th Edition)* Addison Wesley.
- Elmasri, R., Navathe, S.B. (2006): *Fundamentals of Database Systems (5th Edition)*, Addison Wesley; 5 edition.
- Harrison, G. (2015): *Next Generation Databases: NoSQL and Big Data*, Apress, 2015.



## **Part F**

# **Advanced Geoinformatics**

Dr.-Ing. Peter Korduan, Prof. Dr.-Ing. Ralf Bill and M.Sc. Ferdinand Vettermann





# 1 Introduction

One of the emerging technologies in the field of GIS is Internet- or Web-GIS. With the introduction of internet based technologies also new possibilities for the usage of GIS rise. In a first section we give an overview about the technologies used for web-based GIS and the functionality a user can expect. In the second section we describe one popular server program for internet map production and publishing, the **UMN MapServer** (today called MapServer). With this, together with a set of data access, converting and graphic rendering tools, it is simple to set up an individual Internet-GIS application with basic functionality.

On the other hand, the effort to standardise interfaces for interoperable geo-data transfer is necessary to allow many distributed or web-based GIS projects to work. We therefore introduce important norms of the **International Standardisation Organisation (ISO)** and specifications of the **Open Geospatial Consortium (OGC)** in the third section. The last two sections deal with current research topics and recent developments related to spatial data in the internet.

## 2 Internet-GIS

### 2.1 Introduction

With Internet-GIS (often also called WebGIS or Online GIS), spatial and property datasets can be provided to a broad range of users. Based on the technology of the internet, data can be used in small company area networks (intranet), in large world-wide information systems (internet) as well as within the mobile domain through the use of location-based services and mobile GIS (see Part B). More and more information can be integrated through the constant capturing and storing of spatial information, making it available via the internet. However, a simple definition of what constitutes Internet-GIS is hardly possible, since the various specifications and products have been produced by different free committees and application vendors, resulting in a broad range of individual applications of varying quality. Internet technology has been adapted in such a way that nearly all functionalities of a GIS may be delivered in the WWW, but not all are implemented in user-specific applications. Internet-GIS focus on the task of information delivery through web browsers. At the same time, it is becoming possible to access different data sets, from different data providers and made available using different products, using common access standards over the internet.

### 2.2 Terms and range of applications

The principal difference between a GIS and an Internet-GIS is that in the latter case the data is made available over the internet and multiple users have access to the same data at the same time. While a GIS can be a stand alone solution, i.e. only one program on a computer, Internet-GIS is always a **client-server solution**. Nevertheless, the term Internet-GIS is understood with different meanings, ranges of application and functionality. Other terms such as **Online-GIS**, **Web-GIS**, **Web-Maps** or **MapServer** are also often equated with the term Internet-GIS. There is no consensus on the use of this term either with respect to functionality or application. In order to identify and classify the various potential solutions, a schema relating to functionality and technology may be of assistance.

Depending upon the range of applications of the Internet-GIS, a different set of functions is needed. Simple functionality such as interactive mapping (zoom and pan) with spatial queries of the available data and a visual overlay of the information can be regarded as the technological standard. The following broad classes of application can be identified:

- **Simple Information and Query Systems**, which present interactive maps with different themes, together with the relevant attribute data in a standard internet environment (browser) with or without extensions (plug-ins). These are useful within local government e.g. for the publication of information via the internet, especially for the citizen and visitors.
- **Specialized Geo-Information Systems**, which make additional services available either generally or only for a reduced group of users. Functionalities may include address detection, routing or simple data analysis.
- **Web based GI-Clients** with access to a central resource, which provide a number of extended functionalities such as measuring distances, analysis and intersection of the data, export, changing attributes, diagrams etc. Such systems are usually used in an Intranet or Extranet environment as they require support/training and administration of users. Such clients may be used in public authorities as a more economical alternative to proprietary GI-Viewers and Desktop-GIS, as well as allowing data to be easily transferred between locations, e.g. to mobile stations.

- **Geo-data Portals:** Internet applications to provide and/or sell large data sets. The different datasets are usually provided from different geo-data servers. Such applications may be run by local authorities e.g. for the delivery and sale of cadastral data.

A crucial criterion for the choice and development of Internet-GIS is the available bandwidth, and thus the quantity of data which can be transferred. The optimum would be a fast connection line and little data, but in reality the opposite is often the case. It is therefore necessary to balance the needs and desires of users with the available resources for financing maintenance, hardware purchase and software development. An additional problem is that the requirements of the users will tend to change and increase once a system is operational, and it may not be possible to implement the newly-required functionality within the existing technological framework.

For the estimation of the effort to supply certain functionality the following aspects have to be considered:

- What do the users want to do with the Internet-GIS?
- Which groups of users will exist (intranet/internet/extranet)?
- What bandwidth is available (network parameters)?
- Which extensions are necessary for clients and servers (plug-ins, script or programming support)?
- Must additional license costs be paid, for either client or server software, or is there a Free/Open Source Software (FOSS) alternative?
- Which web browsers and operating systems may be used?
- Which data formats are available and/or have to be used?
- How much expenditure can be expected for the development, service and maintenance of the system?

Since the requirements usually increase as the system is used, the designed system should be flexible enough to fulfil current demands and to be extended in future.

### 2.3 Internet-GIS technologies

The base of Internet-GIS is of course the internet itself, which is why client-server technology is also the base for internet-driven GIS. **Client-server technology**, by which geo-data and functions are made available in Internet-GIS, has the following main characteristics:

- The speed depends on the quantity of the data which will be transported as well as on the access mechanism.
- The quantity depends again on the data type.
- The load of the data processing can be distributed on client and servers.
- Client-server technology offers multi-user ability for read, and possibly for write access.

In the range of internet applications there are different types of servers. Each type is specialised for specific tasks. Starting with the **web server**, which is delivering web pages and handles the requests from the clients, there is a long list of other specialised server types. Table 11 lists some of the server types which are relevant for Internet-GIS. Often more than one server is running on the same computer machine. A common combination is a web server working together with a **map server** and a **database server** on one (or several) computer(s) (Figure 151). But while the web and database servers may work in the internet without any further assistance, the map server often has no web interface integrated; it depends on a running web server. Note that most of the servers managing spatial data have standardised interfaces (see section 4) and have reserved names related to their tasks, e.g. the Web Map Server (WMS). These are however not server types, but **service types** in relation to the service they do.

Depending upon the requirements imposed by the data (extent, quality, topicality), and the required functionality of the application, different expenditure for the development is necessary. One has to consider the costs of the software, hardware and their maintenance as well as of the qualified technical personnel. Simple information systems can already be realised with little effort. The simplest form is the representation of static maps in HTML-pages, which are connected with the relevant data simply by clicking in the map graphic (an image).

Server type	Description
Map Server	Deliver raster maps and related thematic information, specialized for users needs. The Map Server renders the map images and can send additionally a client including the map and functionality to interoperate with the map. A Map Server can also work as a web service.
Feature Server	Deliver vector data especially in XML format including the geometry and attribute data of feature objects.
Database Server	Provide thematic data. Can send spatial data too, if spatial extension for the data base management system is available. Connection over ODBC/JDBC.
File Server	Provide files for downloads. Such files containing images, metadata, geo-data can be downloaded in different formats or common office documents.
E-Commerce Server	Store charge models and handle order and obtain operations for geospatial data sets and products.
Application Server	Executes programs for geo-data processing (analysis) and deliver the results.
Terminal Server	Assume all computational work in a client server connection. Exchange only screen data, keyboard input or mouse events with clients.
Service Registry Server	Register web services, e.g. WMS, WFS, ... and enable the retrieval for it. Including Metadata it is called a Metadata Server.
Authentication Server	Handle authentication of users and may know different method for that.
Authorization Server	Store information about what registered user can have access for. User will be authorized related to spatial, thematic and temporal constraints.

Table 11: Important server types related to Internet-GIS

The next step up is the realisation of a map server, also named **web mapping** (Herrmann & Asche, 2001). Here the maps are generated upon request on the server-side. **Scripting-languages** such as ASP, PHP or CGI are used to provide this functionality. More advanced applications with functionality on the client-side may be developed using Javascript, Java, GeoJSON and/or SVG, relatively low-priced or free supplementary product lines. A pre-condition is an existing GIS, from which the data must usually be pre-processed. Therefore the data presented in the Internet-GIS may not be the 'live'/real data, but a copy.

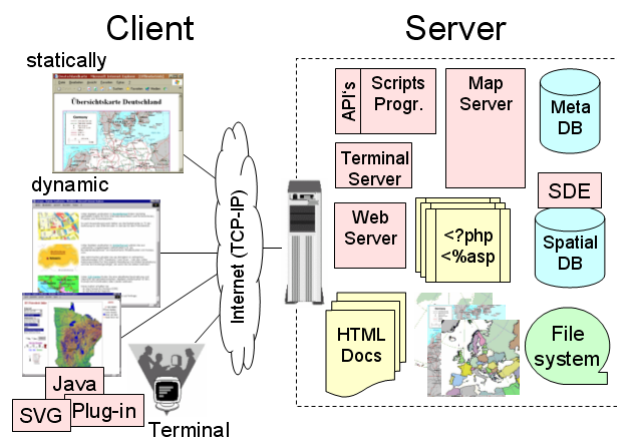


Figure 150: Overview of architectures for Internet-GIS

There are different possibilities for the physical storage of the spatial information. Often one dataset exists for each thematic layer and each region, stored in groups of files in the file system. Larger single datasets, perhaps including entire spatial datasets, require less administrative effort on the file system level but tend to increase the access time required to extract the data. A more sophisticated solution is to use a database with a geospatial extension like the popular PostGIS-extension for PostgreSQL<sup>30</sup>. Such databases allow the combination of attribute and spatial data,

<sup>30</sup> <http://www.postgresql.org>

many different possibilities for indexing, sorting and querying as well as geospatial analysis functions, making them the recommended storage method for spatial data.

An interactive map server with a central geo-data server as backbone becomes necessary if current information from spatial analyses has to be visualised online. Here one can find monolithic proprietary architectures which contain the geo-data servers and the necessary map server as part of a complete GI application. A nearly unrestricted number of Internet-GIS clients can be constructed based on this map server philosophy offering the appropriate functionality to the users. An overview about technology used for web mapping can be found in Seip, Zehner, Korduan, (2017) or Mitchell (2005).

On the client side, more dynamics can be realised with JavaScript and SVG. The principals and advantage of SVG (Scalable Vector Graphics) is described as follows, cited from the W3C:

**Definition:** “SVG is an XML format for describing two-dimensional vector graphics, both static and animated. SVG allows three types of graphic objects:

- vector graphic shapes (e.g. paths consisting of straight lines and curves, and areas bounded by them)
- raster graphics images/digital images
- text.

Graphical objects can be grouped, styled, transformed and composed into previously rendered objects. Text can be in any XML namespace suitable to the application, which enhances search ability and accessibility of the SVG graphics. The feature set includes nested transformations, clipping paths, alpha masks, filter effects, template objects and extensibility. SVG drawings can be dynamic and interactive. The **Document Object Model (DOM)** for SVG, which includes the full XML DOM, allows straightforward and efficient vector graphics animation via ECMAScript or SMIL. A rich set of event handlers such as `onmouseover` and `onclick` can be assigned to any SVG graphical object. Because of its compatibility and leveraging of other Web standards, features like scripting can be done on SVG elements and other XML elements from different namespaces simultaneously within the same web page.”

Using SVG makes it possible to develop a product that is platform independent, something that is crucial to achieve broad acceptance.

Furthermore, another approach is to render maps clientside with **GeoJSON**. This format, developed in 2008, allows to store geographical information in a simple readable format and can be able to provide any kind of vector information. Because that it is based on the **JavaScript Object Notation (JSON)**, it is very easy to integrate and to use in combination with JavaScript. Because of its simplicity many GIS and mapping systems like OpenLayers, Leaflet, MapServer or Geoserver supporting GeoJSON.

## 2.4 Functionality of an Internet-GIS

For the demarcation of typical functionality of an Internet-GIS, the following five broad groups can be identified:

- Static maps
- Dynamic maps
- Editors
- Complete Internet-GIS
- Data management services

**Static maps** are often presented as so-called **clickable maps**. Through the simple mechanism of linking different pre-prepared maps, perhaps at different scales, quite useful results may be achieved. Such a solution delivers the linkage of thematic data with simple navigation/visualisation functionality such as zoom and pan. A special form of static map is the so-called **tiled map**, where a larger map area is divided into individual tiles, which are then linked together by hyperlinks. Both solutions present raster data, which is statically prepared in advance and can not be adapted to the demands of users on-the fly. Tools for the generation and publication of clickable maps are widely available for many standard GI systems. An example is the HTML-ImageMapper from Alta4, allowing the production of reasonably useful solutions. This kind of internet map is more or less the norm, but because of its limited functionality and dynamics it should not be called Internet-GIS. Everything the user sees has to be physically presented and generated before. The expenditure for the care, maintenance, and extension of such a system is accordingly large.

**Dynamic maps** represent the next stage of development. These are characterised by the fact that the maps are generated from the data based on a query from a client. This data may consist of vector or raster data with the results usually being delivered in raster form, although solutions which generate vector data dynamically e.g. in SVG or GeoJSON format, are available. The following functionality is commonly found in such systems:

- Reference/overview map

- Zoom and pan
- Queries from thematic data and map elements
- Object search based on thematic data queries
- Export of graphics
- Measure in the map
- Display of position and end-to-end measurement

Technically, the dynamic maps are often produced using a ‘**map server**’ application. Such a solution is ideally suited for use as an information desk system because of its excellent speed and minimum requirements at the client side. If updating of the data has to be done, or more extensive GIS analysis functions added, a more advanced form of Internet-GIS must be used or the map server solution must be extended, requiring additional programming effort. Using integrated scripting languages, full-functional Internet-GIS can be developed from most map servers.

**UMN MapServer** is a typical map server, providing standard functionality as well as being extensible either through a scripting interface or modification/extension of the source code (see section 3). Some examples of functions which go beyond that which is usually offered by map servers are:

- Spatial query by polygon
- Measurement of area sizes
- Print function
- Export of attribute information and geo-data
- Next neighbourhood search
- Map annotation
- Thematic classification
- Routing/Shortest path
- 3D visualisation

For geospatial analysis functions, special libraries are available such as GeoTools<sup>31</sup> or GEOS<sup>32</sup>.

With regard to data update via Internet-GIS, if only attribute data is to be changed and no graphic display is necessary, e.g. for the change of postal addresses or the updating of cadastral data (ownership records), simple editor functions for databases, using SQL, are sufficient. These are realised usually on top of **web database clients** such as phpMyAdmin<sup>33</sup> for MySQL<sup>34</sup> or pgAdminIV<sup>35</sup> for PostgreSQL or self-developed HTML form elements. For adding or editing of geometry, graphical tools have to be used. Therefore either special plug-ins, JavaScript or Java are used on the client side.

Since **Geography Markup Language (GML)**, (see section 4.3.5) and other standardised vector data formats such as Google KML (see section 6.2.1) or GeoJSON for JavaScript are becoming more popular online, more and more clients provide graphic editor functions. Ready-to-use **application programming interfaces (API)** in JavaScript make it easy to implement simple functions for drawing and modifying points, lines or areas with aerial or satellite images in the background.

**Internet-GIS** with the complete set of GIS functions running in a browser are on-the-way. The expenditure for reprogramming all GIS functionality, especially analysis functions, in JavaScript or Java is high. On the other hand, limitations for applications running in browsers on the client side have to be taken into consideration. Therefore so called **Desktop-Internet-GIS** are beginning to establish themselves on the GIS market. These Internet-GIS are simple desktop-GIS clients, running locally, and can use all resources from the computer, including reading local geo-data, but additionally have an internet connection to access various distributed geo-data resources over the internet. Some applications also focus on the implementation of functions provided by **web processing services** over standardised interfaces such Simple Object Access Protocol (SOAP) or Open Geospatial Consortium Web Processing Service (OGC-WPS), e.g. User-Friendly Desktop Internet GIS (uDIG<sup>36</sup>). Also, the new ArcGIS Pro combines ESRI online services with the classical Desktop GIS ArcMap. Considering this, we are not far away from **fully-functional Internet-GIS**.

Due to its special position, and superficial similarity to fully-functional Internet-GIS, the **terminal server** solution should be briefly mentioned here. In this case, a stand alone GIS, running on the server, is controlled remotely by

---

<sup>31</sup> <http://www.geotools.org>

<sup>32</sup> <http://trac.osgeo.org/geos/>

<sup>33</sup> <http://www.phpmyadmin.net>

<sup>34</sup> <http://www.mysql.com>

<sup>35</sup> <http://www.pgadmin.org>

<sup>36</sup> <http://udig.refractive.net>

the client and thus no new development of GIS functionality or internet compatibility is needed. Well known examples of this architecture are Windows Terminal Server and Citrix.

Finally we need some **data management functions** to organise and maintain geospatial data on the server side. These functions must not be available for all users on the client side and must also not necessarily be realised in graphical interfaces. They are often established to control quality, completeness and to support collaborative work with other users. Functions for geo-data retrieval, subscriptions, and online shop solutions belong to this last group representing data management services.

- Overview of spatial, thematic and temporal properties of the data sets as well as the available functionality.
- Data retrieval making use of a **thesaurus** and **gazetteer**.
- Export of **metadata** related to the indicated view, layer or object in the representation.
- Generalization of metadata, geo-data and thematic data sets for different levels of detail and different users.
- User-specific options for delivery of large volume of data.
- Subscription functions and newsletter.
- Wide area cross-linking.

There are several platforms which are providing these services. The most common ones are the Open Source projects GeoNetwork<sup>37</sup> and CKAN<sup>38</sup> with their spatial extension<sup>39</sup>.

### 3 UMN MapServer

UMN MapServer<sup>40</sup> is an Open Source development environment for building spatially-enabled internet applications. It was originally developed by the University of Minnesota (UMN) ForNet project in cooperation with NASA and others. The server program creates and delivers maps related to certain requirements and interacts using HTTP GET or POST with HTML forms, Flash or Java-applets. MapServer, its naming today, comes with the MapScript API for many programming languages such as PHP, Perl, .Net, Python, Ruby, Tcl or Java. The server works mainly as a CGI application with a web server like Apache. MapServer runs on various operating systems, for example on Linux, Windows or Mac OS X. The standard client runs on any graphical browser without the need for any particular extensions. MapServer clients can also be implemented using SVG or Flash and may therefor need plug-ins.

Geographic data are typically sourced from conventional GIS formats, such as ESRI shapefiles or MapInfo coverages (Figure 151). All vector and raster formats supported by the GDAL/OGR<sup>41</sup> library can be read. Geo-data can also be sourced from one or more spatial database tables (e.g. PostGIS, SDE, Oracle or MySQL). The databases can be distributed over the network. The server can include geo-data from distributed WMS and WFS as well as serve data in a number of formats, following open OGC standards (e.g. WFS, WMS, WCS, SOS). MapServer uses free configured thematic and/or spatial filter, styles, symbols and fonts for user specific visualization. Vector data can be classified for thematic maps. MapServer supports SLD for WMS layer.

#### 3.1 MapServer CGI

MapServer runs normally as a CGI (Common Gateway Interface) application, controlled with CGI-parameters. The parameter mode determines how MapServer processes the query. The default mode is BROWSE, where MapServer creates the maps and builds the interactive client page with template. For the dynamic modification of the template MapServer uses substitution strings, marked with square brackets. All other modes create parts of the client content or do single navigation tasks. The mode MAP creates only the map and can be used in conjunction with the HTML `<img ...>` tag. Other parts can be created separately with the modes REFERENCE, SCALEBAR and LEGEND. ZOOMIN and ZOOMOUT do exactly this, but with a specified zoom factor and positive or negative ZOOMDIR and switch to the mode BROWSE. The mode QUERY is used for spatial search closest to a point clicked previously in a map. NQUERY finds more than the nearest feature from the selected search point or user-defined search box. A text search is also available with ITEMQUERY and ITEMNQUERY. Modes for querying features are also available with MapServer CGI. Kropla (2005) provides a good reference for using MapServer as a CGI application to maps for the internet.

---

<sup>37</sup> <https://geonetwork-opensource.org/>

<sup>38</sup> <https://ckan.org/>

<sup>39</sup> <https://github.com/ckan/ckanext-spatial>

<sup>40</sup> <http://mapserver.gis.umn.edu>

<sup>41</sup> <http://www.gdal.org>

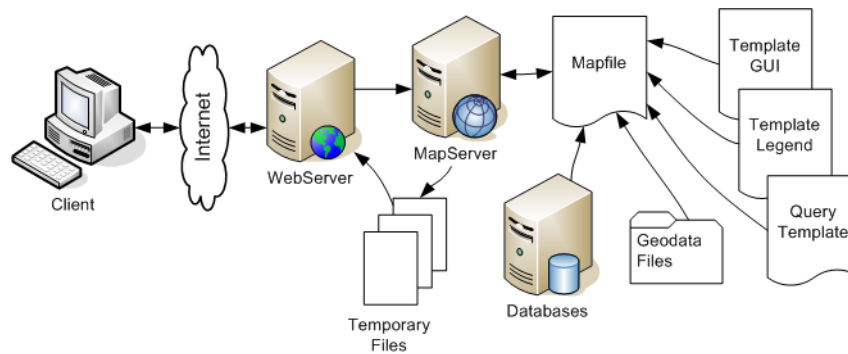


Figure 151: Components of MapServer

## 3.2 Mapfile

The mapfile is the central configuration file for MapServer. It includes the description of the data resources that are to be generated for display in the client and define the layout of it. More than one data source, in the GIS context often called layer, can be combined in a mapfile. The file structure is divided into different sections, also called objects, because these sections are represented as map objects in the MapScript API, described in section 3.3.

### 3.2.1 Map section

The all enclosing section is the **MAP** section. A section starts with the name of the section as a keyword, indicated here with upper-case letters, and finishes with the keyword **END**. Each section can have parameters represented in the map objects by attributes. Values may be of different types: constants, strings or numbers. Variable names in the mapfile are not case-sensitive, but attribute names or paths are. For a better overview, keywords are written here in upper-case italics. File paths can be given absolute or relative to the mapfile. Paths to data sources can also be set relative to the path assigned to the keyword *SHAPEPATH*. The first declarations in the **MAP** sections relate to general settings of the map, e.g. a *NAME* for the map, the map images *SIZE*, the spatial *EXTENT* the map will cover, *UNITS* and path where the application should look for symbol (*SYMBOLSET*) and font (*FONTSET*) definitions.

### 3.2.2 Web section

The **WEB** section includes declarations of where the temporary files are stored (*IMAGEPATH*) on the server and the relative location of the temp directory from which the client can get the results over the HTTP connection (*IMAGEURL*). This section also includes the definition of the main *TEMPLATE*. The template is normally a HTML page containing placeholders for the dynamically changing parts such as the map, legend, scale bar etc. Other contents of this section are the path to files for logging (*LOG*) and error messages (*ERROR*) and metadata describing the map (*METADATA*). The metadata are in a separate section ending and consist of content information, contact information and other things related to the map, such as the supported spatial reference system(s). The metadata will also be useful if MapServer should run as OGC Web Service (see section 4). The **REFERENCEMAP**, **LEGEND** and the **SCALEBAR** are also separate sections. Legends can be created by MapServer as a single image or more dynamically with a template containing form elements for group, layer, class and query control. Otherwise only the size of the legend symbols and the font can be varied. A **QUERYMAP** can be included in a query template to display the objects found in a map.

### 3.2.3 Projection section

The **PROJECTION** section is important to describe the projection and the spatial reference system (SRS) which MapServer shall use for the planar map presentation. The declaration can be a projection code from the European Petrol Survey Group (EPSG) database or user-defined projection values. The section can also be used for layers, described next, to define the projection of the underlying local or remote geo-data sources. If no projection section is used, all layers will be displayed in the same system as they are defined in the data source and no projection will be applied. The map extent must be given in the same SRS as defined in the projection section.

### 3.2.4 Layer section

The **LAYER** section describes the source and the style how and when to display the data. *NAME* and *STATUS* should be given for all layer declarations. The variable *GROUP* makes it possible to group layers together, but for the time being only one level of grouping is supported by MapServer. With the variable *REQUIRES*, constraints for displaying a layer in relation to the status of other layers can be set. For example one background raster layer can automatically be switched ON if a special vector layer is set to the *STATUS ON*. *MINSCALEDENOM* and *MAXSCALEDENOM* are useful to change the visibility of layers in the client depending on the map scale.



Depending on the layer TYPE, point, line, polygon, circle, annotation, raster or query, some of the required variables change. For vector data, the variable *CONNECTIONTYPE* defines the type of the geo-data the layer comes from. The *Default* is local and MapServer assumes it must read ESRI shapefiles. If OGR is chosen, all vector data formats supported by the OGR library can be used as a layer source. Connection types *postgis*, *oraclespatial*, *sde* and *mysql* are available to connect to geospatial databases. The *CONNECTION* variable specifies the connection string for a database connection. *DATA* defines the shapefile or raster file name, or the identification of the source depending on the OGR format. The layer section also includes a *TEMPLATE* variable that points to the file used as template for the query result sets. In the case that more than one feature is found with a query in the NQUERY mode of MapServer, the template will be used n times for each result set and has to be combined with a *HEADER* and *FOOTER* template to build a complete HTML page. For queries, table joins can also be used in the layer section. The *JOIN* section defines in this case the table and identifier to join with the thematic table of the data resource. Vector layer uses *CLASS* sections to fulfil thematic classification or restrict the displayed content by attributive constraints, defined by the *EXPRESSION* variable. A spatial restriction is possible using the *FILTER* variable within the LAYER section or, if the layer is from a database, by WHERE clauses in the SQL string of the *DATA* variable (see Part E). The graphic representation of a *CLASS* can be defined with a single *COLOR*, a *SYMBOL* or a definition of one or more *STYLE* sections. Other variables define the *SIZE*, *WIDTH*, *OUTLINECOLOR* and other style properties. The combination of styles and symbols determines the graphical layout of the map. There are additional sections and variables to configure the map layout, including background colour map grind, labels and annotations.

### 3.3 MapServer extensions

The most comprehensive and valuable extension of MapServer is the **MapScript API**. The original version of MapScript was written in Perl and uses SWIG<sup>42</sup>, but since SWIG does not support the PHP language, the module has to be maintained separately and may not always be in sync with the Perl version. The MapScript API exists for several other programming languages, such as Python, C#, Python, Java or Microsoft .net. However, one of the most widely-used script languages in the internet is PHP. The PHP module phpMapScript was developed and is currently maintained by DM Solutions Group<sup>43</sup>. The MapScript API is composed of map objects similar to the elements of the mapfile, together with methods for manipulation and interacting with the objects. Due to the fact that nearly all parameters specified in the mapfile are properties of the corresponding objects in MapScript, functions for the creation of map components can be called separately. With the capabilities of the programming languages, map objects can be build on persistent data that must not necessarily be included in a mapfile. Map components can therefore also be composed from data stored in a database or other resources (see section 5.1.3 and Figure 152).

This offers a great deal of flexibility to developers. A map object function can for example also be used without any configuration data or with an empty map object. phpMapScript belongs to the MapServer source code and can be compiled with the option `--with-php`. In the following example the `php_mapscript` library is loaded, in the case MapServer is running on a Windows OS or on another platform (e.g. Linux). The next step is the instantiation of the map object based on the mapfile `mymapfile.map`. With the function `draw()` the temporary map will be drawn. The result is the name of the temporary file, which can be included as source in an Image HTML tag, delivered to the client through the Web Server.

```
<?php
dl('php_mapscript.so');
$map_path="/var/www/html/ms/map_files/";
$map=ms_newMapObj($map_path." mymapfile.map");
$image=$map->draw();
?>
<HTML>
  <HEAD>
    <TITLE>Example Map </TITLE>
  </HEAD>
  <BODY>
    
  </BODY>
</HTML>
```

<sup>42</sup> <http://www.swig.org>

<sup>43</sup> <http://www.dmsolutions.ca>

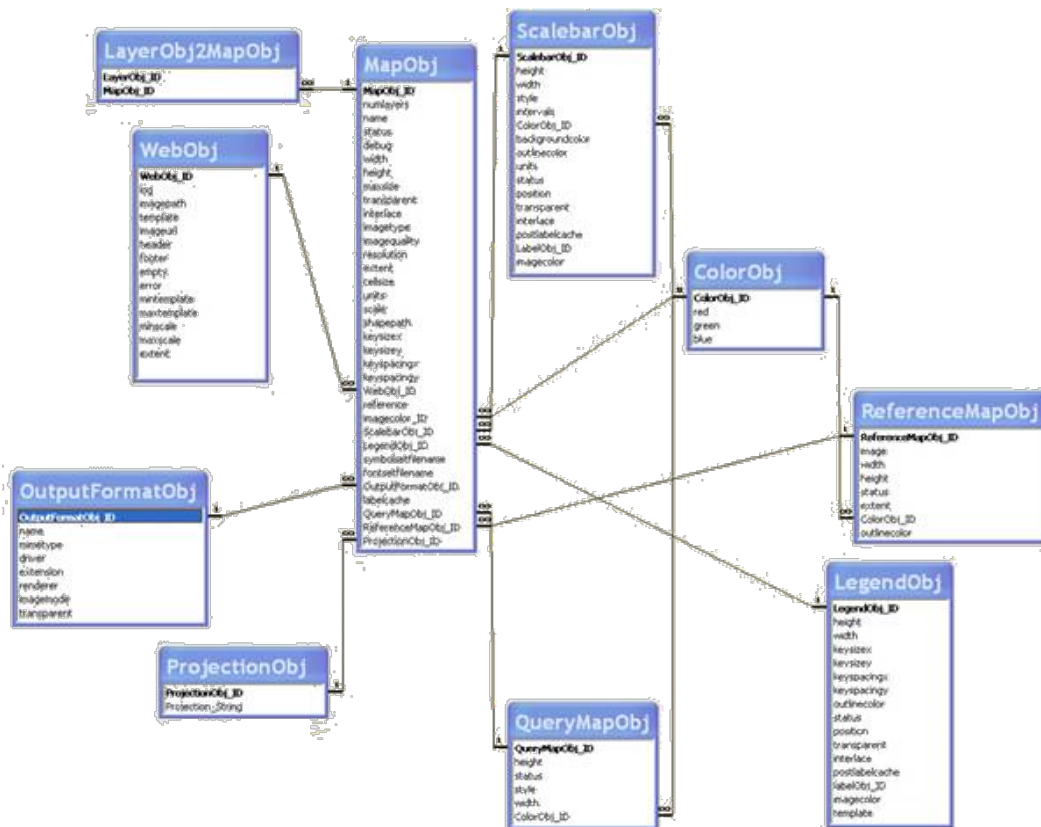


Figure 152: Relational database model for MapServer map objects

### 3.4 MapServer Clients

The basic MapServer software package comes without a standard client, because MapServer is a server application. But there are several MapServer packages, client examples and tutorials that help developers quickly develop their own Internet-GIS application based on MapServer. Two examples with precompiled MapServer applications and several different clients for MapServer developed both from DM-Solution are ms4w for Windows and FGS for Linux. Additional to the base module containing Apache web server with PHP, MapServer, phpMapScript, library for geo-data access and database connection support, many client packages with example data are available, e.g. Gmap and ms\_ogc\_workshop<sup>44</sup>.



Figure 153: A basic HTML MapServer client

Derived from the different internet technology solutions, we can classify MapServer clients into different types. There are simple HTML clients, only equipped with form elements for user interaction with the map, legend, layer selection and other parts of the client interface. A more comfortable client is equipped with JavaScript functions

<sup>44</sup> <http://mapserver.github.io/ms-ogc-workshop/>

to support more dynamic client side interaction. For MapServer also Flash, SVG and Java clients exist. A new form of smart mapping client uses so-called AJAX functions, see sections 2 and 6.2.4.

## 4 OGC and ISO - interoperability and standardisation initiatives

### 4.1 Introduction

The widespread application of computers and use of geographic information systems (GIS) have led to the increased analysis of geographic data within multiple disciplines. Based on advances in information technology, society's reliance on such data is growing. Geographic datasets are increasingly being shared, exchanged, and used for purposes other than their producers' intended one. GIS, remote sensing, automated mapping and facilities management (AM/FM), traffic analysis, navigation systems, and other technologies for Geographic Information (GI) entered a period of radical integration.

The dissemination of the spatial information is based on web services in the internet. Standards for the spatial information description are defined and published by **ISO (International Organisation for Standardization)**. Abstract and implementation specifications are established by the **Open Geospatial Consortium<sup>45</sup> (OGC)** and implemented in many GIS products on the market. Since 1997, ISO and OGC have been working closely together in TOCG (TC211 – OGC coordination group). The major outcomes of this process are described in the following sections.

Beside ISO and OGC there are other organisations defining the most common standards, e.g. the **World Wide Web Consortium<sup>46</sup> (W3C)** with XML (Extensible Markup Language), DTD (Document Type Definition), XSL (Extensible Stylesheet Language) and SVG (Scalable Vector Graphics). Others are the **Object Management Group<sup>47</sup> (OMG)** and the **IEEE<sup>48</sup>**.

An international standard provides a framework for developers to create software that enables users to access and process geographic data from a variety of sources across a generic computing interface within an open information technology environment.

- “a framework for developers” means that an international standard is based on a comprehensive, common plan for interoperable geoprocessing.
- “access and process” means that geo-data users can query remote databases and control remote processing resources, and also take advantage of other distributed computing technologies such as software delivered to the user's local environment from a remote environment for temporary use.
- “from a variety of sources” means that users will have access to data acquired in a variety of ways and stored in a wide variety of relational and non-relational databases.
- “across a generic computing interface” means that ISO 19119 interfaces provide reliable communication between otherwise disparate software resources that are equipped to use these interfaces.
- “within an open information technology environment” means that an international standard enables geoprocessing to take place outside of the closed environment of monolithic GIS, remote sensing, and AM/FM systems that control and restrict database, user interface, network, and data manipulation functions.

### 4.2 ISO standards

ISO is the union of national standardisation institutions. Since November 1994 a technical committee (**TC 211<sup>49</sup>**, another one is TC 204 dealing with car navigation) has been developing standards in the field of digital geographic information. “This work aims to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth. These standards may specify, for geographic information, methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations. The work shall link to appropriate standards for information technology and data where possible, and provide a framework for the development of sector-specific applications using geographic data. The overall objectives of ISO/TC 211 are:

- to increase the understanding and usage of geographic information;
- to increase the availability, access, integration, and sharing of geographic information;

---

<sup>45</sup> <http://www.opengeospatial.org>

<sup>46</sup> <http://www.w3c.org>

<sup>47</sup> <http://www.omg.org>

<sup>48</sup> <http://www.ieee.org>

<sup>49</sup> <http://www.isotc211.org>

- to promote the efficient, effective, and economic use of digital geographic information and associated hardware and software systems;
- to contribute to a unified approach to addressing global ecological and humanitarian problems.” (<http://www.isotc211.org/>)

TC 211 has around 40 participating members and more than 30 observing member countries and many external liaisons with working groups and project teams from scientific and other communities. Currently there are 6 working groups, for example Geospatial services or Imagery.

ISO TC 211 has defined and published around 77 standards til now, a selection of the ISO 191xx series is shown in Table 12 (recent list from <https://www.iso.org/committee/54904/x/catalogue/p/1/u/0/w/0/d/0>). For further details see Kresse & Fadaie, 2004.

ISO Standard	Title
ISO 6709:2008	Standard representation of latitude, longitude and altitude for geographic point locations
ISO 19101-1:2014	Geographic information - Reference model - Fundamentals
ISO 19101-2:2018	Geographic information - Reference model - Imagery
ISO/TS 19103:2005	Geographic information - Conceptual schema language
ISO 19107:2003	Geographic information - Spatial schema
ISO 19108:2002	Geographic information - Temporal schema (Cor 1:2006)
ISO 19109:2015	Geographic information - Rules for application schema
ISO 19110:2016	Geographic information - Methodology for feature cataloguing
ISO 19111:2007	Geographic information - Spatial referencing by coordinates
ISO 19112:2003	Geographic information - Spatial referencing by geographic identifiers
ISO 19113:2002	Geographic information - Quality principles
ISO 19114:2003	Geographic information - Quality evaluation procedures (Cor 1:2005)
ISO 19115:2003	Geographic information - Metadata (Cor 1:2006)
ISO 19116:2004	Geographic information - Positioning services
ISO 19117:2012	Geographic information - Portrayal
ISO 19118:20011	Geographic information - Encoding
ISO 19119:2016	Geographic information - Services
ISO/TR 19120:2001	Geographic information - Functional standards
ISO/TR 19121:2000	Geographic information - Imagery and gridded data
ISO/TR 19122:2004	Geographic information / Geomatics - Qualification and certification of personnel
ISO 19123:2005	Geographic information - Schema for coverage geometry and functions
ISO 19125-1:2004	Geographic information - Simple feature access - Part 1: Common architecture
ISO 19125-2:2004	Geographic information - Simple feature access - Part 2: SQL option
ISO/TS 19127:2005	Geographic information - Geodetic codes and parameters
ISO 19128:2005	Geographic information - Web map server interface
ISO 19133:2005	Geographic information - Location-based services - Tracking and navigation
ISO 19135:2005	Geographic information - Procedures for item registration

Table 12: Selected ISO standards for geographic information produced by ISO/TC211

### 4.3 Open Geospatial Consortium standards

“The Open Geospatial Consortium, Inc.® (OGC) is a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services.” (<http://www.opengeospatial.org/>). The OGC was founded in 1994 and originally called the Open GIS Consortium. The OGC integrates more than 500 members (GIS-vendors, service providers, IT- and DB-enterprises, data suppliers, universities and others). Different grades of membership reflect the influence in the standardisation process and the financial engagement in OGC: strategic membership, principal membership, technical membership and many other types of associate membership. Since 1999 the OGC has had a strong focus on internet-based solutions. The main goal is **interoperability**.

**Definition:** Interoperability describes a technology, which allows application developers and users to use all kind of geocoded data and geo-functionality or -process being available in the net inside his own environment and individual workflows.

OpenGIS® Specifications are the main “products” of the OGC. These are technical documents that detail interfaces or encodings, which software developers may use to build support for the interfaces or encodings into their products and services. These specifications address specific interoperability challenges. Ideally, when specifications are implemented by two different software engineers working independently, the resulting components plug and play, that is they work together without further debugging. The documents are available at no cost to everyone. The “**Abstract Specification**” provides the conceptual foundation for most OGC specification development activities. Open interfaces and protocols are built and referenced against the Abstract Specification, thus enabling interoperability between different brands and different kinds of spatial processing systems. At the present time OGC has defined around 20 abstract specifications. The Abstract Specification provides a reference model for the development of OpenGIS Implementation Specifications. “**Implementation Specifications**” are different from the Abstract Specification. They are written for a more technical audience and detail the interface structure between software components. An interface specification is considered to be at the implementation level of detail, and, when implemented by two different software engineers in ignorance of each other, the resulting components plug and play with each other at that interface.

Many implementation specifications have currently been agreed (version numbers in brackets), named are the most prominent ones:

- *Web Map Service (WMS 1.3.0)* - creation and display of registered and superimposed map-like views of information as raster maps/imagery
- *Web Feature Service (WFS 2.0.2)* - to retrieve and update geospatial data encoded in vector form, typically as Geography Markup Language (GML), from multiple Web Feature Services
- *Web Coverage Service (WCS 2.1)* - extends the Web Map Service (WMS) interface to allow access to geospatial coverages (raster data sets) that represent values or properties of geographic locations.
- *Catalogue Service Implementation Specification (CSW 3.0)* - defines a common interface that enables diverse but conformant applications to perform discovery, browse and query operations against distributed heterogeneous catalogue servers.
- *Coordinate Transformation Service Implementation Specification (1.0)* - provides interfaces for general positioning, coordinate systems, and coordinate transformations.
- *Filter Encoding Implementation Specification (FE 2.02)* - defines an XML encoding for filter expressions.
- *Geographic Objects Implementation Specification (GO 1.0.0)* - defines an open set of common, lightweight, language-independent abstractions for describing, managing, rendering, and manipulating geometric and geographic objects within an application programming environment.
- *Geography Markup Language (GML 3.3)* - is an XML encoding for the transport and storage of geographic information, including both the spatial and non-spatial properties of geographic features.
- *Grid Coverage Service Implementation Specification (2.1)* - defines methods that allow interoperability between software implementations by data vendors and software vendors providing grid (raster) analysis and processing capabilities.
- *Location Service (OpenLS) Implementation Specification (1.2)* - is an open platform for location-based application services
- *Simple Features Implementation Specifications 1 and 2 (1.2.1), CORBA (1.0), OLE/COM (1.1) and SQL (1.2.1)* - define interfaces that enable transparent access to geographic data held in heterogeneous processing systems on distributed computing platforms.

- *Styled Layer Descriptor Implementation Specification (SLD 1.1.0)* - is an encoding that extends the Web Map Service specification to allow user-defined symbolization of feature data. It allows users (or other systems) to determine which features or layers are rendered with which colours or symbols.
- *Web Map Context Implementation Specification (WMC 1.1)* - a companion to the OpenGIS® Web Map Service describing how to save a map view comprised of many different layers from different Web Map Servers. A 'context' can be encoded and saved so that Web maps created by users can be automatically reconstructed and augmented by the authoring user or other users in the future.
- *Web Service Common Implementation Specification (2.0.0)* - details many of the aspects that are, or will be, common to all OGC Web Service interface Implementation Specifications.

All together they define the base for a distributed, heterogeneous GI-architecture. Almost all vendors of GIS products produce software based on one or more of these standards.

The OGC defines abstract and implementation specifications. An example of an abstract specification is the OpenGIS Service Architecture specification in which architecture patterns for services used for geographic information are described, and a services taxonomy and list of examples are defined. The Service Architecture Specification (Topic 12 of OGC Abstract Specifications) describes how to create platform-neutral service specifications. An example for an implementation specification is the Web Map Service Implementation Specification, which describes interfaces and operations for the exchange of rendered geographic data (maps).

Terms relating to the abstract definition of geospatial interoperability are service and operation.

- **Service:** A collection of operations, accessible through an interface, that allows a user to evoke a behaviour of value to the user.
- **Operation:** specification of an interaction that can be requested from an object to effect behaviour.

Terms relating to the implementation of geospatial interoperability are interface and component:

- **Interface:** an implementation of operations including the syntax of the interaction for a given distributed computing technology.
- **Component:** a physical, replaceable part of a system that packages implementation and conforms to and provides the realization of a set of interfaces. Component is synonymous with Server.

The definition for service, operation and interface are taken from ISO 19119.

#### 4.3.1 Web Map Service (WMS)

The *Web Map Service* supports delivery of raster data (images) from geo-referenced data. The maps are supplied in common image formats. The WMS specification defines three operations.

- *GetCapabilities* sends metadata about the service's content, the supported operations and the accepted request parameter.
- *GetMap* responds with maps (images) in accordance to the parameters sent with the request.
- *GetFeatureInfo* supplies thematic information of a feature displayed in the map and selected through its position.

The last of these operations is optional. That means the service must not implement this operation: not all geographic datasets offer thematic information (e.g. a raster dataset).

#### GetCapabilities

Each *OGC Web Service (OWS)* has at least a *GetCapabilities* operation, to inform the client about the service capabilities. The specification of the operation is such that the request is identical for all services, but a different response is given depending on the service type: each service type has different capabilities metadata. The parameters of the *GetCapabilities* operation are shown in Table 13.

Parameter	Value	Meaning
Service	e.g. WMS, WFS	Which service is called
Request	GetCapabilities	Which operation shall be invoked
Version	e.g. 1.3.0	Which version of the service is supported

Table 13: GetCapabilities request parameters

A *GetCapabilities* Request may look like:

<http://www.ows-server.org/owsprog?SERVICE=WMS&VERSION=1.3.0&REQUEST=GetCapabilities>

The result of every *GetCapabilities* request is an XML document. Conformant to the OGC service architecture abstract specification, the information included in the XML document must be sufficient to understand the meaning

of the offered layers and to build valid requests to obtain the underlying geographic data sets. All information needed for further interaction with the service must be included in the GetCapabilities document. In some cases special operations are offered to get more detailed information, e.g. about layers, feature types or observation behaviours.

In the case of a WMS a GetCapabilities response consists of a service and a capabilities section. The service section includes the service metadata such as name, title, abstracts, keywords and online resource but also contact information about the organisation which provides and/or published the data with the service and other details such as use and access constraints. As a minimum of descriptive information the mandatory metadata elements have to be set as specified in ISO 19119. Optional and vendor specific elements can also be included depending on local metadata profiles and the requirements of specialised service catalogues.

The capabilities section includes as a minimum the sub-sections request, exception and layer. In the request section we can find the supported operations of a service. As previously mentioned, the number and type of operations depends on the type of the service. WMS supports GetCapabilities and GetMap as a minimum and optionally GetFeatureInfo. If Styled Layer Descriptor (SLD) is supported by this WMS (available since version 1.3.0), the operations DescribeLayer, GetLegendGraphic and GetStyles may be found in the request section. In the exception section details for behaviour in the case of errors on the server side or invalid request parameters or values sent from the client side are given. All services must provide at least an XML formatted error message. Optionally the message may be send in plain-text format or, in the case of WMS, within an otherwise empty image in the requested image format. It is in the responsibility of the client software to handle error messages.

```
<WMT_MS_Capabilities version="1.1.1">
- <Service>
  <Name>OGC:WMS</Name>
  <Title>Vietnam-Demo</Title>
  <Abstract>Vietnam geospatial demo data based on shape files</Abstract>
  <OnlineResource xlink:href="http://localhost/cgi-bin/germany_demo.cgi?"/>
+ <ContactInformation></ContactInformation>
  <AccessConstraints>none</AccessConstraints>
</Service>
- <Capability>
- <Request>
  + <GetCapabilities></GetCapabilities>
  + <GetMap></GetMap>
  + <GetFeatureInfo></GetFeatureInfo>
  + <DescribeLayer></DescribeLayer>
  + <GetLegendGraphic></GetLegendGraphic>
  + <GetStyles></GetStyles>
</Request>
+ <Exception></Exception>
  <VendorSpecificCapabilities/>
  <UserDefinedSymbolization SupportSLD="1" UserLayer="0" UserStyle="1" RemoteWFS="0"/>
+ <Layer></Layer>
</Capability>
</WMT_MS_Capabilities>
```

Figure 154: Part of a capabilities document with WMS service metadata

The layer section can be subdivided in a hierarchical structure of any depth, but many clients support up to only two levels, meaning that layers can only be grouped at one level. Only rarely are sub-level layer groups supported. Layer metadata are included comparable with the service metadata: common descriptive information about the content of the layer and additional information how to use the layer. Beside a minimum of a title and a name of the layer, the supported spatial reference system as well as bounding box, style and scale hints can be given.

## GetMap

Within the capabilities document all required information for the generic production of a valid and useful GetMap request are available. The request string is built from the description of the OnlineResource element. Whilst the GetCapabilities operation must be implemented with the HTTP GET request method, the GetMap operation can be called using either GET or POST. In the following example we show a typical GET request:

```
http://a-map-co.com/mapserver.cgi?VERSION=1.1.1&REQUEST=GetMap
&SRS=EPSG:4326&BBOX=-97.105,24.913,78.794,36.358&WIDTH=560
&HEIGHT=350&LAYERS=AVHRR-09-27&STYLES=&FORMAT=image/png
&BGCOLOR=0xFFFFFFFF&TRANSPARENT=TRUE
&EXCEPTIONS=application/vnd.ogc.se_inimage
```

The parameters used in a GetMap request are described in Table 14.

Request parameter	Mandatory / Optional	Description
VERSION=version	M	Request version
REQUEST=GetMap	M	Request name
LAYERS=layer_list	M	Comma-separated list of one or more map layers. Optional if SLD parameter is present
STYLES=style_list	M	Comma-separated list of one rendering style per requested layer. Optional if SLD parameter is present
CRS=namespace:identifier	M	Spatial Reference System
BBOX=minx,miny,maxx,maxy	M	Bounding box corners (lower left, upper right) in SRS units
WIDTH=output_width	M	Width in pixels of map picture
HEIGHT=output_height	M	Height in pixels of map picture
FORMAT=output_format	M	Output format of map
TRANSPARENT=TRUE/FALSE	O	Background transparency of map (default=FALSE)
BGCOLOR=color_value	O	Hexadecimal red-green-blue colour value for the background colour (default=0xFFFFFF).
EXCEPTIONS=exception_format	O	The format in which exceptions are to be reported by the WMS (default=SE_XML)
TIME=time	O	Time value of layer desired
ELEVATION=elevation	O	Elevation of layer desired
Other sample dimension(s)	O	Value of other dimensions as appropriate
Vendor-specific parameters	O	Optional experimental parameters
SLD=styledlayer_descriptor_URL	O	URL of Styled Layer Descriptor (as defined in SLD Specification). Only with SLD supporting WMS
WFS=web_feature_service_URL	O	URL of Web Feature Service providing features to be symbolized using SLD. Only with SLD-WMS

Table 14: WMS 1.1.1 GetMap request parameters

With the selection of sufficient parameters, the user has the ability to query for raster data corresponding to their demands. Transparent vector layers can be overlaid with other layers. The user needs to only use the same data format, image size, extent, and spatial reference system. It is a standard solution to overlay different layers from distributed sources and realise an added value. The WMS is widespread because of its simple implementation with GET requests on the client side. A link to a geo-data resource can be integrated in a simple way as the source for an HTML image

```

```

### GetFeatureInfo

A further operation of the WMS is specified to query the service for thematic information related to the objects displayed in the previously produced map. In a GetFeatureInfo request the user sends the position in pixel coordinates (X and Y) for which they want to retrieve information. Usually this is the position where the user has clicked with the mouse. So that the WMS knows to what the X and Y pixel position is related to, the GetFeatureInfo query must include the GetMap query parameters of the image the position is related to. Normally these are the parameters of the latest GetMap request. The WMS delivers information on one or all of the layers selected for the query. The client tells the server for which layer information should be retrieved using the parameter QUERY\_LAYERS. The optional parameter INFO\_FORMAT is used to choose the format (MIME type) of feature information. Only the formats given in the capabilities document are supported. Another optional parameter, FEATURE\_COUNT can be used to limit the number of returned features. The GetFeatureInfo request has no ability for complex, expression-like queries. Sometimes vendor specific parameters are used, e.g. to define a search



radius around the click point. A final optional parameter which can be added to all requests is EXCEPTIONS that define the MIME type to be used for error messages or other messages describing exceptions. Here again, only the types that are declared in the capabilities document are supported.

### 4.3.2 Styled Layer Descriptor (SLD)

The Styled Layer Descriptor Implementation Specification describes a format for a map-styling language for producing georeferenced maps with user-defined styling, see OGC (2002d). Alternatively the user may choose from a list of styles the WMS provides or accept a default style. A user-defined style can be sent to the WMS by setting an URL reference, using the SLD parameter in a GetMap request or by sending the style definition with the request in the parameter SLD\_BODY. Three additional operations for WMS are available to implement full SLD support. Defining a user-defined style requires knowledge about the features being symbolised, or at least their feature type. To query this information, the DescribeLayer request can be used, e.g.:

```
http://server.com/wms?VERSION=1.1.0&REQUEST=DescribeLayer&LAYERS=Rivers
```

If SLD is supported by the WMS, the capabilities document describes these enhanced functions. If a WMS is to symbolise features using a user-defined symbolisation, it is necessary to identify the source of the feature data. The SLD specification has foreseen that the user can symbolise feature or coverage data stored in a remote Web Feature Service (WFS) or Web Coverage Service (WCS). If this option is supported, the otherwise optional parameters REMOTE\_OWS\_TYPE and REMOTE\_OWS\_URL can be used. These parameters will be sent with the HTTP-GET GetMap request to direct the WMS to a remote WFS or WCS service as the 'default' source for feature/coverage data, e.g.:

```
http://server.com/wms?VERSION=1.3.0&REQUEST=GetMap&
SRS=EPSG:4326&BBOX=9,52,15,58&
SLD=http://client.com/SLD.xml&WIDTH=300&HEIGHT=300&FORMAT=PNG&
REMOTE_OWS_TYPE=WFS&REMOTE_OWS_URL=http://otherserver.com/wfs?...
```

Other additional optional operations for SLD-WMS are DescribeLegendGraphic, GetStyle and PutStyle. With the first one of these, images can be queried that depict the symbols related to the styles used. The final two can be used to get or respectively to put a style description from or to a list of styles stored on a server. With PutStyle, new styles can be inserted or updated. The styles are described in XML documents with special symbolisers for the different geometry types, text, labels and raster graphic.

### 4.3.3 Web Feature Service (WFS)

WFS enables feature-level access to spatial data and vector data exchange using Geography Markup Language (GML). WFS describes a rich interface for spatial, thematic and temporal queries and can be optionally implemented with transactional capability, called then Web Feature Service - Transactional (WFS-T). The GetCapabilities and DescribeFeatureType operations provide information about the service, the queryable features and the available query capabilities. With GetFeature, the user has direct access to the vector geometry and the associated thematic information. If transactions are supported, the operation may be extended to GetFeatureWithLock. The locking can also be realised without getting features with the operation LockFeature. The operation Transaction is implemented to enable a client to send newly created or modified features back to the server or to delete features. The parameter OPERATION defines how the server should handle the sent feature. Possible options are INSERT, UPDATE and DELETE. The latter operation can be sent with a GET request. The others must be sent with HTTP-POST because in these cases geometry must be attached to the request. This could result in long GET strings which may cause problems with the web server or intervening servers handling the request. If transactions are supported by a WFS, security aspects have to be taken into consideration. UMN MapServer (described in section 3) supports only the basic WFS with no transactional support.

#### GetFeature

The GetFeature operation allows users to retrieve feature geo-data. The canonical representation of features uses GML. With the parameter OutputFormat the user can specify in which format the requested features should be sent, e.g. "text/gml; subtype=gml/3.1.1". All features are modelled as belonging to a feature type. With the operation GetCapabilities the user can query for a list of supported feature types. With DescribeFeatureType a detailed description of feature types is queryable. If the client is informed about supported feature types it can request features types within the GetFeature by naming feature type names in the parameter typeName, separated using a comma. Another means to restrict the result set is the use of filters. With filter statements, single features can be queried according to the properties set in the filter. A simple filter is to select features by identification numbers, e.g.:

```
<Query typeName="myns:InWaterA_1M">
  <ogc:Filter>
    <ogc:GmlObjectId gml:id="InWaterA_1M.1013"/>
```

```
<ogc:GmlObjectId gml:id="InWaterA_1M.1014"/>
<ogc:GmlObjectId gml:id="InWaterA_1M.1015"/>
</ogc:Filter>
</Query>
```

Filter statements in a POST request are contained within a Query tag. A filter in a GET request looks like this:

```
typename=FEAT1,FEAT2&filter=(<Filter>... FEAT1 filter...</Filter>)(<Filter>...
FEAT2 filter...</Filter>)
```

#### 4.3.4 Web Coverage Service (WCS)

The WCS is the raster equivalent to the WFS described above. In version 2.1.0, the service is not limited to regular grids anymore as it was in earlier versions. Usually a WCS provides real raster data such as a digital elevation model (DEM) or a GeoTIFF with multiple bands e.g. for satellite imagery (see Part C of this textbook). The available operations are GetCapabilities, DescribeCoverage and GetCoverage. With these, simple query functions with spatial and temporal filter capabilities are supported. GetCapabilities is comparable with all other OGC web services. DescribeCoverage delivers a complete description of one or multiple coverages called with its identifiers separated by commas.

The GetCoverage operation has, together with the usual service, request and version parameters, the required parameters identifier, to identify the requested coverage, and format. Bounding box, time sequence and other settings for the response grid can also be set in the request.

#### 4.3.5 Geography Markup Language (GML)

“In GML a feature is represented as an XML element. The name of the feature element indicates the Feature Type, conventionally given in UpperCamelCase, such as `xmml:BoreHole` or `myns:SecondaryCollege`. The content of a feature element is a set of elements, which describes the feature in terms of a set of properties. Each child element of the feature element is a property. The name of the property element indicates the property type, conventionally given in LowerCamelCase, such as `gml:boundedBy` or `xmml:collarLocation`. The value of a property is given in-line by the content of the property element, or by-reference as the value of a resource identified in a link carried as an XML attribute of the property element. If the in-line form is used, then the content may be a literal (a number, text, etc), or may be structured using XML elements, but no assumptions can be made about the structure of the value of a property. In some cases the value of a property of a feature may be another feature, for example a `myns:School` feature may have a property `myns:frontsOn`, whose value is a `myns:Road`, which will itself have further properties, etc. However, note that the properties of the second feature (the `myns:Road`) are not properties of the first feature (the `myns:School`) and it is an error to refer to them as such.” (OGC, 2005a, OGC 2005b)

### 4.4 Metadata standards

Geo-data are more complex than other data due to its spatial behaviour and relation. It is more complicated to understand the data and to identify the correct meaning of the content. If one reads for example a raster map, such as a satellite image, it is not immediately obvious where the object depicted in the image is located, unless the user has prior knowledge about the scene. Therefore geo-data needs a special spatial description. The thematic and temporal aspects of a data set also have to be described for users that don't have foreknowledge of the data, but intend to use it or are searching for it. Another reason to describe geo data explicitly is the effort required by a user to open the data and explore the content. If, for example, a user searches in a large data collection for a special type of geo-data, they have to use a GIS program in order to open the data and view it. It is not reasonable to expect users to carry out data retrieval with GIS software and to have to open each data set. Indeed, it may not be possible to open all data sets during a search operation as some may only be available to purchase.

In simple terms, meta-information is “The information and documentation which makes data sets understandable and shareable for users” (ISO 11179). In the context of ISO 19115, metadata are “Data describing the content, quality, condition, format, and/or other characteristics of more basic data”. The metadata standard recommends the use of core elements which may be mandatory or conditional (see Figure 155).

Mandatory Elements:	Conditional Elements:
Dataset title	Dataset responsible party
Dataset reference date	Geographic location by coordinates
Dataset language	Dataset character set
Dataset topic category	Spatial resolution
Abstract	Distribution format
Metadata point of contact	Spatial representation type
Metadata date stamp	Reference system
	Lineage statement
	On-line Resource
	Metadata file identifier
	Metadata standard name
	Metadata standard version
	Metadata language
	Metadata character set

Figure 155: ISO19115 core metadata elements

To address this problem we require data about data, so called **metadata**. Metadata are human- and machine-readable data which are separated from the data to be described.

- Metadata enables effective, efficient, and accurate use of data sets and data collections.
- Metadata is the starting point for every kind of discovery system.
- Metadata is a documentation of collected data, understandable many years after its collection.
- Metadata is the base for interoperability (data exchange).
- Metadata allows different detailed views on the whole data set.
- Metadata presents the production processes externally.
- Metadata supports external data exchange and data management between customers, service companies and administration.

To make data retrieval possible for a large and distributed set of spatial information and to make metadata comparable they must be standardized. The ISO and OGC have specified a metadata standard for the description of spatial information. In ISO nomenclature the standard "Geographic information – Metadata" has the number 19115.

Originally the standard was derived from the Content Standard for Digital Geospatial Metadata (CSDGM) from the US Federal Geographic Data Committee (FGDC). The FGDC started in 1994 with a national spatial data initiative and the standardisation process for metadata. The OGC took the CSDGM and developed an object-oriented standard with the focus on modern UML-driven modelling techniques. The content of the ISO 19115 is identical to the Topic 11: Metadata in the Abstract Specification of the OGC. It is also a good example of the collaborative work between ISO and the OGC. Any information community may create their own profile of ISO 19115, with the requirement that it includes the 13 core element.

## 5 Application example: Internet-GIS for municipalities

### 5.1.1 Introduction

A large part of the data held in local government has spatial relations, both text data such as addresses or street kilometres (so-called secondary metrics, Bill, 2016) as well as coordinate-based data (so-called primary metrics, Bill, 2016) such as cadastral- or topographic maps. Nearly all municipalities have severe financial problems and therefore an increasing requirement to make cost savings. By using the spatial references to make a meaningful linkage between the geographical base data and the different technical and administrative datasets, the access and the development of these may be facilitated substantially. A new term has been adopted for such a linkage of eGovernment and GIS technologies: **‘GeoGovernment’**. The increased accessibility of the datasets resulting from the use of Internet-GIS technology means that this technology is becoming increasingly important for local government. Workloads can be reduced and operating times can be shortened, and thus resources can be saved. From the combination and intersection of the different information levels a new quality of data and service levels may result. The local agenda process puts additional pressure on the administrative offices, because of its requirement to inform citizens about social, ecological and economical facts in the municipality or county. Here the use of internet technology and Free/Open Source Software (FOSS) offers a low-cost but functionally advanced solution.

Client-server based Internet-GIS technology offers the possibility that the data remains, and can therefore be administered and maintained, at the place where it is produced. Additionally, the data processing and analysis

programs can run on the server, meaning that the user of the data does not need to worry about specific software requirements on their computer – they need only to run a standard web browser on a thin client and can achieve their goals by simply querying the data and visualising the results. This saves much time, which would be necessary in learning to use complex software, as well as for the data management and also reduces the costs of hardware and software. In addition, the responsibility for the data is still in the hand of the data producer and provider. However, they then have to prepare a user-friendly human interface to their complex data, usually necessitating a major development effort for the individual governmental offices. This burden can be lessened by making use of FOSS which supports easier transfer of developments made by other software developers or administrative offices, and often results in solutions which fit better to the needs of users.

In order to facilitate the publication of the authority's data on the internet, some initial preparation is necessary with regards to the formats and storage methods of the data. It is a legal requirement in Germany that the cadastral office of each municipality or county has to set up a digital cadastral data set. For this process, a more or less unique approach is foreseen for the whole of Germany, named ALKIS (Authoritative Real Estate and Cadastral System). The graphical data set is given with the real estate map; this handles the spatial and graphical data relating to parcels, buildings and other geodetic information, usually at a scale between 1:1.000 and 1:2.500. The tabular digital property data is the automated real estate book; this includes ownership, parcel size, land usage and other legal rights related to parcels and buildings. The content and the structure of this ALKIS GIS solutions was finished in the last years for whole Germany. The legislation also requires that other administrative spatial information systems, such as environmental information systems or planning information systems (see Bill, 2016), should be based on these so-called geo base data sets. Local administrations, which acquire and update these base data sets, are therefore increasingly searching for solutions in order to publish and deliver these data to other users in as easy and straightforward manner as possible. Cadastral offices are placing themselves in the role of data service providers. This should create positive impacts on the geo-data market in general. In the district administration, many items of data have a common spatial reference. The linkage of these items should result in a new quality for the access to the data. It will also improve the data management, e.g. by reducing redundancies.

### 5.1.2 Data collection and conversion

In the first phase of the realisation of the Internet-GIS for the county administration, all information which should be published via the internet needed to be converted to a suitable format. This applied to ALKIS and other data sets on the county level. The official transfer format of ALKIS is NAS (Norm-based exchange format).

For more flexibility in data access and linking, as well as to provide support for a larger volume of users, the district administration now may develop its own database structure. The standard output and transfer format is calling WLDGE. These plain-text files must then be converted into SQL-statements to create and update the database tables. The developed converter, written in PHP, is a separate server-side software component, supporting both MySQL and PostgreSQL databases and is published as `wldge2sql` under the GNU Public License (GPL)<sup>50</sup>. With these conversion steps it is possible to hold both cadastral spatial information and data about the landowner in one database. This leads to added value of the data.

The implemented solution should give different users access to the common data base, both on the county level and the municipality level. Additional data sets might be integrated, e.g. at the county level environmental data, and at the district level zoning plans. In many municipalities the zoning plans had been only available in an analogue format, i.e. it was necessary to scan and georeference these maps – the map server therefore delivers these maps as raster data. In addition, one may integrate digital orthophotos or high precision satellite imagery which are nowadays available in many countries.

Each new application needs the input of new and more specific data. For the purposes of land register management, a large number of documents were digitised by scanning. Online forms were used to provide metadata, which was uploaded together with the appropriate files via HTTP to the server database. Calculation of geothermal energy potential needs specific data for the soil and ground water levels. For this, the position and attribute information will be stored together in the geospatial database. Currently many data sets have to be specifically prepared and converted for publication via the internet. As data volumes increase, this will become increasingly problematic. However, through the introduction of a better technical infrastructure at the various administrative offices using interoperable middleware conforming to the previously mentioned OGC WMS/WFS standards (see section 4.3), this conversion process should become obsolete.

---

<sup>50</sup> <http://www.gnu.org/licenses/gpl.html>

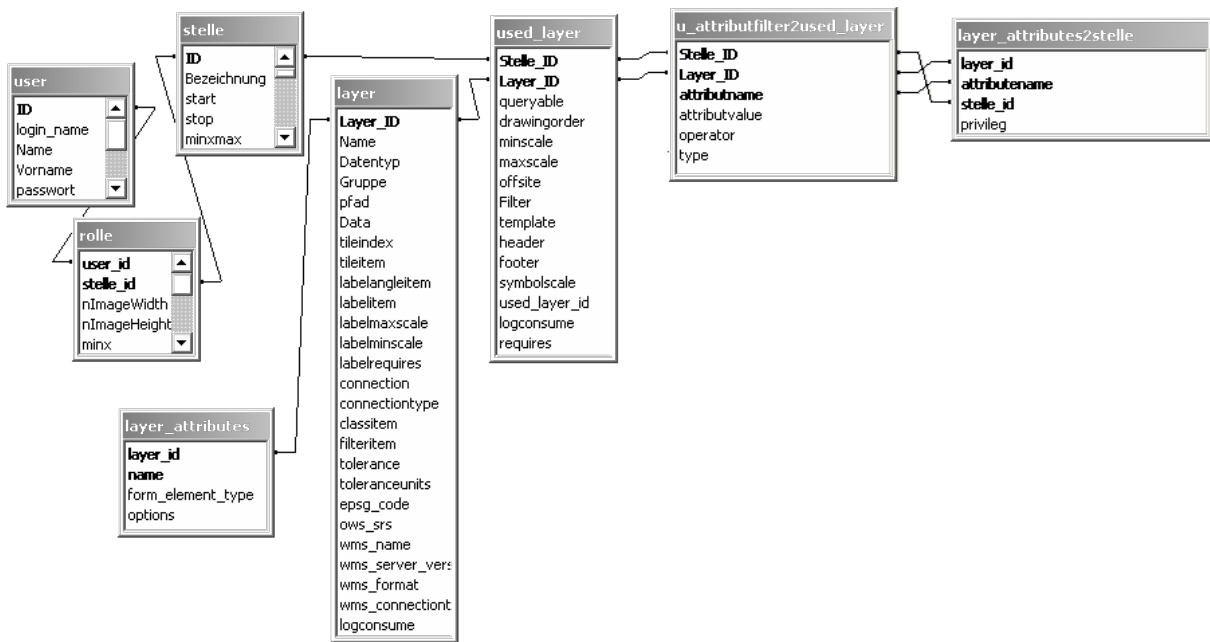


Figure 156: Database model for users, positions and roles

### 5.1.3 System architecture

The system can be used by multiple users at the same time. This is inherent in the client-server architecture, with communication via the HTTP protocol. A user can use the system in a specific role. After client authentication the user must choose a position in which they will work and the system then provides the specific interface for this role. User, role and position have appropriate properties in the data model, allowing relevant permissions to be assigned. A use case for the usage of the system is shown in Figure 157. The user “official” has more access rights. They can use a cadastral information desk whereas the normal “user” only can use the common information desk. Special use cases and tasks for employees of a municipality, e.g. the possibility to input data for land register management, were assigned by position rather than by user. This recognises the fact that a user may have different tasks, and the tasks may change for a user, but are relatively persistent for a position. Additionally, a person may work in more than one position. The interface properties for a user are task-independent (e.g. screen size), whereas the properties of a position are related to the tasks a user can do in that position (e.g. which data is presented to them). The exact interface for a role is dependent on the user selecting the role and their position (e.g. last minimum enclosing rectangle of a map and the visible layer). Thus, multiple users may carry out the same role in the same position but have personalised interfaces. The layout of the map, usually configured in single mapfiles, is also stored in a database. To manage this data, a relational database model was developed, (Figure 156).

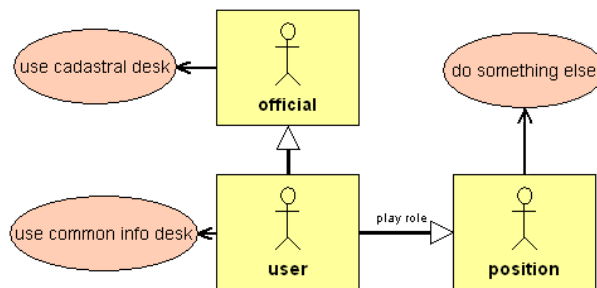


Figure 157: Actor specification to carry out a specific role for a position

All actions within the system were modelled as use cases. The client sends a request to the server, which contains the name of the use case (e.g. searching for land parcels) and some parameter to qualify the request. The server responds with a web page containing a map, a form or a specific formatted document (e.g. a PDF-document with information about land parcel ownership). The load of the data processing is distributed between client and server. Most of the processing will be done by the server, notably the map requests will be fulfilled by the server scripts. Only in the cases where more interaction is necessary is the client equipped with more processing functionality (e.g. distance measuring, digitising spatial objects, etc.).

### 5.1.4 Technical realisation and results

The implementation of the Internet-GIS took place on the basis of UMN MapServer (see section 3). The human interface was realised using HTML, Java-applets and SVG, with some specific menu functionality programmed in JavaScript. The MapServer map files for configuration have been replaced by a dynamic configuration using the phpMapScript map object and database project tables (see Figure 152 and Figure 156). All configuration options, e.g. which role a user may occupy, which layers to present and which menu bars may appear, were stored in the tables and related to a user, role or position. The internet application starts for a new user with an overview map of the entire county of Bad Doberan, see Figure 158. Besides the administrative borders of the municipalities, further data sets to be used on the county level are integrated here. Typical map scales presented here are in the range of 1:2.500 to 1:50.000, with topographic raster maps used as a backdrop. Thematic data in this view are, for instance, maps of contaminated sites or maps of natural monuments. This Internet-GIS solution can then become a standard for querying interdisciplinary data sets on the county level. This could reduce redundancies, make information flows transparent and avoid the need for transfer of paper and digital copies between the different specialised offices in the county's administration. At this level, data security or data protection are not so critical. Later this level might also be used as a starting point for a citizen information system for the whole county.

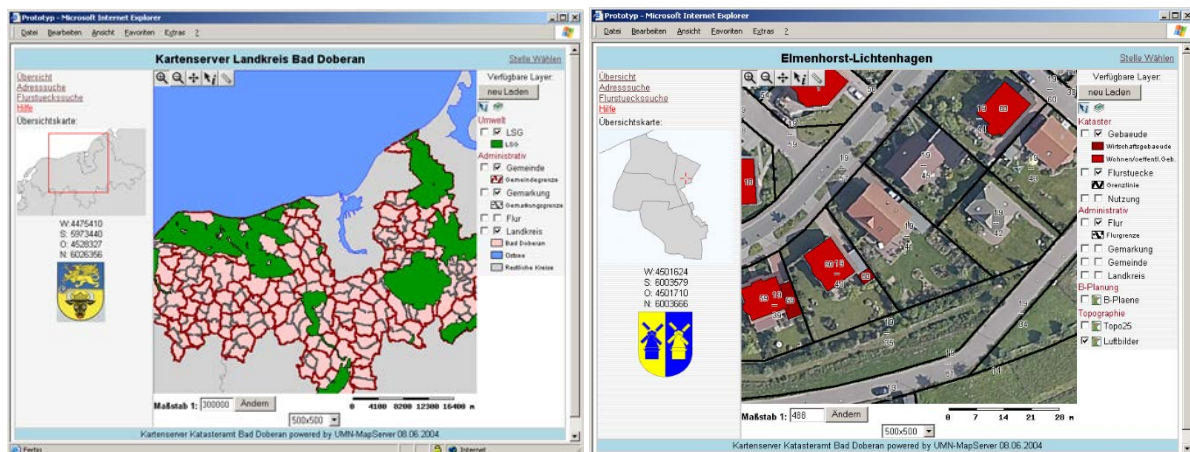


Figure 158: Graphical User Interface (GUI) with access permission on county and municipality level

Participating communes can start from this county level to reach the municipality level. At this level, data protection is increasingly important. The authentication with username and password can be realised using standard HTTP authorisation through the Apache Web Server. Only users with the appropriate permissions can reach this protected level. This fulfils the legal data protection requirements, e.g. cadastral data on ownership may by law only be viewed by the owner or appropriate officials in the administrative offices. Within the protected range for municipal employees, again the Internet-GIS starts with an overview map of the commune showing municipality borders and further administrative subdivisions of the commune such as field borders. For special users at municipality level different views may be generated. Each position can be configured in its spatial, thematic and temporal extension, depending upon what a user in a role is granted to see and do as well as what is relevant for a specific task.

In the cadastral shell the land register data **ALKIS** (parcels and buildings, land ownership) is offered. Simply by querying the ALKIS data of individual parcels or buildings at the appropriate zoom level the user retrieves the information on the ownership of parcels. These query results are prepared in a format of exactly the same layout and design the user would see when working with analogue data.

With such an Internet-GIS solution one can fulfil more or less all requirements an official user at the municipality level would ask for, i.e. the local administration officer no longer needs a stand-alone GIS simply for querying ownership information from the cadastral office. Without any additional costs (hardware/software) or additional effort (data conversion, training etc.) they are now able to query the cadastral data using a standard web browser. This is a major benefit for the municipality exchequer.

In the topographic section we may integrate **ATKIS** or raster topographic maps at a scale of 1:25.000 and georeferenced aerial photographs with a ground resolution of approximately 15cm. In many use cases, especially for planning and topographical purposes, the high resolution data set may replace the official cadastral map.

In the planning section the **generic zoning plan** (scale 1:5.000 to 1:25.000 dependent on the size of the communes) and the detailed **individual zoning plans** of the municipalities, usually in a scale of 1:2.500 to 1:1.000, are simply visualised by their border and a text placement, if not available in GIS form. The scanned and georeferenced detailed zoning plan is integrated here as a raster data set and is selected for visualisation by just clicking on the text annotation. If this plan is available in vector format it could replace the raster data. In addition, the textual part of the detailed zoning plan, the plan notation and explanation of the signatures etc. are presented as text documents.

Beside these three realisations for different users on a municipality level, other sections (environment, technical infrastructure etc.) may easily be integrated with the existing digital data. The additional page elements (borders, logos, etc.) at municipality level can be produced in a completely different layout than on the county level. This gives the municipalities the chance to bring their own corporate identity within the webpage design. It also allows the communes to create different internet services on top of these data sets, such as links to the administrative offices, advertisements and so on. For this, a uniform internet appearance within the municipality is desirable.

### 5.1.5 From mapping to GIS

For more interactive functions the user must have the possibility to insert or modify data. These tasks place a greater requirement on the client browser and the data exchange mechanism. A multi-user data access is easy to realise, but, locking and transaction mechanisms are essential for data writing. To provide more functionality on the client side, we introduced Scalable Vector Graphic (SVG) in the application (see section 2.3). Using SVG makes it possible to develop a product that is platform-independent, something that is crucial to achieve broad acceptance. We introduced SVG for the first three use cases of the Internet-GIS application:

- Calculation of geothermal energy potential and location of drill holes.
- On screen digitising of sealed urban surfaces over aerial ortho-photographs.
- Meta-information system for the management of survey plans, coordinate register and certificates of the land register (geo-referencing and document retrieval).

One example of the district's internal procedures is the survey of **geothermal energy potential** in upper soil regions (near surface, see Figure 159). The district's targets are the development of:

- methods for providing geothermal data,
- optimized equipment for the needs of clients and carriers,
- a new nationwide economy sector for geothermal energy.

In this context, the effects of accruing heat deprivation in adjacent soil regions are of special interest to the district's planning strategy. Another example of the district's internal procedures is to verify the level of sealing of urban surfaces. Providing the sealed surface area is a legal duty for the county. This has to be done in order to calculate taxes dealing with the maintenance of rain runoffs (i.e. wastewater, see Figure 159).

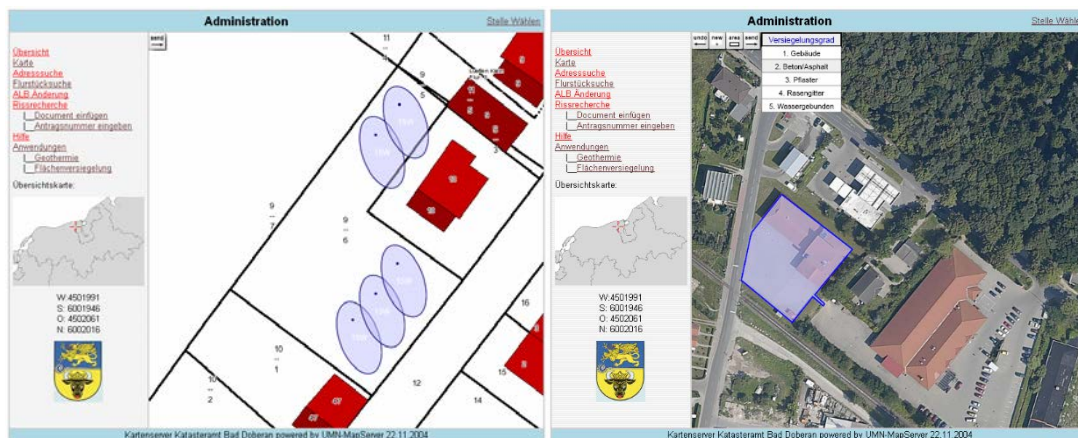


Figure 159: SVG used for drill holes disposition in geo thermal (left) and capturing sealed urban surfaces

The third use case using SVG functionality is that of a meta-information system for the management of land register documents. Meta-information systems play an important role for management of spatial data both in municipalities and city councils (see Korduan, 2003). The client provides on-screen digitisation of polygons, which represent the spatial extent of the content in the documents. The polygons as well as textural metadata about the documents are provided in a form and sent to the server via HTTP. Polygons and text are stored in a PostgreSQL database. The spatial extension of PostgreSQL, PostGIS, provides functionality to select documents by a given search polygon. Documents can therefore be retrieved and collected for the work preparation for survey offices,

## 6 Recent developments and research issues

### 6.1 Sensor Web Enablement (SWE)

A **sensor network** is a computer-accessible network of many spatially distributed devices using sensors to monitor conditions at different locations, such as temperature, sound, vibration, pressure, motion or pollutants. A **Sensor**

**Web** refers to web-accessible sensor networks and archived sensor data that can be discovered and accessed using standard protocols and application programming interfaces (APIs).

The OGC has defined **Sensor Web Enablement (SWE)** which is focused on standards to enable the discovery, exchange and processing of sensor observations as well as the tasking of sensor systems.

“The functionality that OGC has targeted within a sensor web includes:

- Discovery of sensor systems, observations, and observation processes that meet our immediate needs
- Determination of a sensor’s capabilities and quality of measurements
- Access to sensor parameters that automatically allow software to process and geolocate observations
- Retrieval of real-time or time-series observations and coverages in standard encodings
- Tasking of sensors to acquire observations of interest
- Subscription to and publishing of alerts to be issued by sensors or sensor services based upon certain criteria“ (OGC; 2006b)

In the Sensor Web Enablement initiative of the OGC, several standards have been developed, such as:

- **Observations & Measurements Schema (O&M)** – Standard models and XML schema for encoding observations and measurements from a sensor, both archived and real-time.
- **Sensor Model Language (SensorML)** – Standard models and XML schema for describing sensors systems and processes; provides information needed for discovery of sensors, location of sensor observations, processing of low-level sensor observations, and listing of taskable properties.
- **Sensor Observations Service (SOS)** - Standard web service interface for requesting, filtering, and retrieving observations and sensor system information. This is the intermediary between a client and an observation repository or near real-time sensor channel.
- **Sensor Planning Service (SPS)** – Standard web service interface for requesting user-driven acquisitions and observations. This is the intermediary between a client and a sensor collection management environment.
- **Sensor Alert Service (SAS)** – Standard web service interface for publishing and subscribing to alerts from sensors.
- **Web Notification Services (WNS)** – Standard web service interface for asynchronous delivery of messages or alerts from SAS and SPS web services and other elements of service workflows.

The vision of the OGC is to develop a standard to “plug-and-play” web-based sensors. The sensor location is thereby a critical parameter and the standards are harmonised with other OGC specifications. Further relations are to sensor and alerting standards such as the IEEE 1451. Different modern location surveying methods such as GNSS and Cell-ID with triangulation makes mobile sensors capable of reporting their geographic location along with their collected data. XML is used to publish formal descriptions of sensor capabilities, locations and interfaces. This allows web-based sensors to be automatically discovered without prior knowledge.

## 6.2 Earth viewers

### 6.2.1 Google Earth

One of the most famous Earth viewers is Google Earth. It is also known as a planet browser. The client of Google Earth starts up with the visualisation of the Earth as a sphere. If the user zooms in, each point of the Earth’s surface can be explored from a bird’s-eye view. Bird’s-eye views are sourced from satellite images and aerial photographs. The Google Earth client software downloads the required data from a list of available servers and mirror servers. This leads to respectable speed for map display on the screen. Google Earth also supports a 3D view. Users can make rapid interactions with zoom, tilt and rotate as well as fly over in the scene, for example through a canyon represented as a digital 3D model. The satellite images are mapped onto a digital elevation model for this.

Google Earth is developed based on the former Earth viewer of “Keyhole”. This company was taken over in 2004 by Google. Since 2015, Google is distributing all versions for free and partially also as Open Source (Google Earth Enterprise). The Desktop Version of Google Earth (Google Earth Pro) is proclaimed as the ultimate research, presentation and collaboration tool for geo-specific information. Additionally, an Enterprise version allows the combination of enterprise data with Google Earth data delivered as ASP solution. It is also possible for users to host their own complete data set. Google Earth Enterprise includes the parts Fusion, Server and Enterprise Client. The first of these integrates custom data points, vector, terrain data and raster images from various GIS in major formats. The Google Earth Server streams these data to client software such as the Enterprise client or Google Maps in a browser.

Since 2010, Google Earth is integrated in Google Maps. Moreover a browser plugin for chrome as well as a smartphone app is also available. The images shown in the display of the application have appropriate resolutions



depending on the scale. Users can place placemarks, polygons, links, and image overlays over the map and append some thematic information that will appear in a label as a so-called “balloon”.

All created data can be exported as XML files in the **Keyhole Markup Language (KML)** format. Large data collections can be compressed in the format KMZ. The KML format stores not only the data but also the views on the data. So the position of a standpoint together with a view direction and tilt can be stored in such a KML file. Furthermore, KML stores the display styles of the spatial objects and can group and order it in folders. The root element is `<kml>`. `<document>` and `<folder>` follow in a hierarchical structure. These tags are containers for a layer. If layers are grouped in folders, these will be displayed in the sidebar of the Google Earth client when the KML file is loaded. The geometry types Point, LineString, LinearRing, Polygon and MultiGeometry are supported. All points can be defined with 3D coordinates with the latitude and longitude in the World Geodetic System 1984 (WGS84) and the elevation over ground. This allows floating objects to be displayed in the 3D views.

### 6.2.2 Bing Maps

Competition between Google and Microsoft, with the Bing Maps viewer (first published as Virtual Earth), is leading to a rapid development in the sector of Earth viewers. In contrast to Google Earth, Bing Maps supports in its web app natively only a birds eye perspective. For the 3D mode a plug-in is necessary. Furthermore, the 3D extension is only available for the bigger cities. Another difference is that users of Google Earth more keyboard functions, e.g. for panning, and the resolution of the images are often much better in Google Earth than in Bing Maps. The base application for Bing Maps is Microsoft Bing (<https://www.bing.com/maps>). A completely unique feature in Virtual Earth in comparison to other Earth viewers is the ability of viewing an oblique view – the birds eye view (see Part C). The user can choose from high resolution images taken from the four celestial directions at an angle of 45° to the ground. These images are only available for larger cities in some countries.

### 6.2.3 NASA World Wind

An Open Source application in the field of Earth viewers is NASA WorldWind<sup>51</sup>. WorldWind lets you zoom from satellite altitude into any place on Earth, as in Google Earth. Leveraging Landsat satellite imagery and Shuttle Radar Topography Mission data, World Wind allows users to experience the Earth’s terrain in visually rich 3D. The imagery data are sourced from the NASA satellite missions: Landsat 7, SRTM, NASA SVS and MODIS data are available. Other planets, such as Mars and Venus, can be explored with this software. WorldWind comes with a variety of visual guides that enhance the user’s experience, such as latitude and longitude lines, as well as extremely precise coordinate data. These helpers can be toggled on or off at any time and are viewable with any of WorldWind’s other features turned on. Nasa World Wind offers also the possibility to view other planets, like Google Earth. Furthermore, the software has an interface to Java, Android and the integration into HTML as well as a kit for server side developments.

### 6.2.4 Google Maps API

For the development of user-specific internet GIS applications, Google offers an API for its Google Maps service. This is a JavaScript API running exclusively in browsers on the client side. Users can include the API with a key, for which they must register. The key can only apply for one directory, and its subdirectories, on one server. This means that the web pages that include the API can only be stored in one of these directories. The API is included as follows:

```
<script language="JavaScript"
  src="http://maps.google.com/maps?file=api&v=2.58&key=12xyz">
</script>
```

The base of the API version 3 is the *google.maps.Map* Object. The the older versions are deprecated. A documentation of the API and a full API class reference can be found under <https://developers.google.com/maps/documentation/javascript/tutorial>.

In the first step of instantiation of the *Map* object the name of a HTML DIV element is assigned. The DIV element is the place in the HTML page in which the map should be displayed. Additional parameters, such as the size of the map and the behaviour with respect to mouse interactions can be assigned to the *Map* object. The Map object has methods for the configuration of the map, controls, map types, map status, overlays, info windows, projection and events. Many of the values are predefined with default values. In Google Maps, only predefined zoom levels are supported. The images used in Google Maps will not be rendered afresh after each request, unlike in the MapServer solution. All maps are produced as tiles of a size of 256×256 pixels. The images will only be loaded if they are necessary. This demonstrates the use of **AJAX (Asynchronous Javascript And XML)** technology. If the client drags the map, all required new tiles will be downloaded. The API provides *MapOptions* objects for the control of the map view. For example, buttons and a ruler are available e.g. to switch between different map types

---

<sup>51</sup> <http://worldwind.arc.nasa.gov>

(map, satellite, hybrid or user defined map types) or to zoom. The developer can set the desired position of the map controls using the *position* property of the Map object.

The concept of drawing vector features over the raster map is called **overlay**. Overlay is an interface class: all overlays such as *Markers*, *Shapes*, *Info Windows* are derived from the Overlay class. The Icon object can be used to style Markers, the geographic representation of points in the map. InfoWindow is necessary for the presentation of text information. With Custom Popups additional links, images, other HTML elements, and also additional JavaScript code, such as other Google maps, can be displayed.

With the class *Geocoder Service*, geographic coordinates can be queried from given postal addresses or vice versa. With the *Event* object, special events can be registered. These events are often bound to mouse interactions e.g. that an event will be triggered if the mouse leaves the shape of an Overlay object.

The Google Maps API also provides the ability to develop user-specific controls, e.g. to query data or integrate images. The same is possible for overlays. The displayed parts of overlays (icon, shadow, infowindow...) will be composed of different HTML DIV elements. These DIV elements are organised in different MapPane objects and levels with a strict drawing order. The developer has to decide which element in which pane has to be drawn. Last but not least, the API gives the ability to include user-specific spatial data as new map types through the interface class *GTileLayer*. With this class, new map types can be defined and sources linked together. This can also be applied for layers sourced from OGC web services, for instance a WMS layer. A prepared script to enable the easy loading of WMS layers into a user's Google maps application is demonstrated under <http://www.sumbera.com/lab/GoogleV3/tiledWMSoverlayGoogleV3.htm>. In this Google Maps API extension all required WMS parameters are integrated with a Ground Overlay object.

### 6.2.5 OpenLayers

As mentioned above, each user needs a registration key for using Google Maps API, and the free version is limited to only private use. Alternatively an open source smart mapping browser such as the API OpenLayers<sup>52</sup> can be used. This object-based JavaScript API includes comparable objects and functions to the Google Maps API, but has a stronger focus on open geospatial standards such as those specified by the OGC.

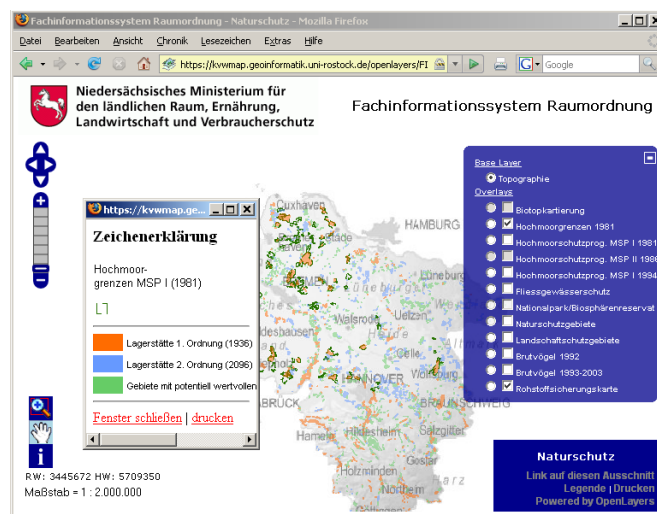


Figure 160: OpenLayers GUI for an area management system

## 6.3 Mashups

A mashup in the spatial context is a web application that combines spatial with non-spatial data from one or more sources into an integrated experience. The etymology of this term derives from its similar use in pop music, possibly from the hip-hop music practice of mixing two or more songs. Typically, the content will be sourced from standardised interfaces or APIs. Interfaces for users to create their own mashups have also been developed. These allow the user to drag and drop data points into the map application, e.g. Google Maps. In this case with kml various kinds of mashups can be created. A common mashup for photographs is Flickr: thousands of photos are geo-referenced and visualised in a Google map. Further examples can be found under <http://www.panoramio.com> or <http://wikimapia.org>. The latter example focuses on the spatial location of the terms described e.g. in Wikipedia.

<sup>52</sup> <http://www.openlayers.org>

## 6.4 GeoRSS

*Geographically Encoded Objects for RSS (GeoRSS)* is a simple XML-based data format that contains information and messages with a spatial location. GeoRSS is based on RSS 2.0. Real Simple Syndication (RSS) is an XML format for transmitting news and the content of news-like sites and personal weblogs. RSS can however be used not only for news, but for anything that can be broken down into discrete items. RSS can therefore also be used to carry spatial information. RSS is becoming more and more prevalent as a way to publish and share information in the internet. The simplest way to extend RSS with a spatial location is to add the tags `<geo:lat>` and `<geo:long>`, resulting in GeoRSS as defined by the Resource Description Framework (RDF) standard. A GeoRSS RDF can be added to each RSS Channel:

```
<?xml version="1.0"?>
<rss version="2.0" xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <channel>
    <item>
      <title>Example</title>
      <link>http://www.geoinformatik.uni-rostock.de/</link>
      <description>News</description>
      <geo:lat>12.3323</geo:lat>
      <geo:long>54.3289</geo:long>
    </item>
  </channel>
</rss>
```

Also defined are GeoRSS Simple and GeoRSS URL. GeoRSS Simple uses `<georss>` tags instead of `<geo:...>` tags:

```
<georss:point>12.3323 54.3289</georss:point>
```

Another form is GeoRSS GML for geometry objects with more complexity:

```
<georss:where>
  <gml:Point>
    <gml:pos>12.3323 54.3289</gml:pos>
  </gml:Point>
</georss:where>
```

All these solutions are used in many web portals with spatial visualisation of information. Leading companies which uses GeoRSS are Google in Google Maps and Google Earth and also Yahoo with Yahoo Maps. Yahoo Maps supports a further GeoRSS specification<sup>53</sup>.

Some API's, e.g. Google Maps API have special classes to handle GeoRSS feeds. In OpenLayers 3+ the GeoRSS must be transformed into compatible formats like GeoJSON. For this there are several libraries available, for example GeoRSSToGeoJSON<sup>54</sup>.

The OGC has also specified a GeoRSS version to harmonise the different implementations of a spatial extension to RSS.

## References

- Bill, R. (2016): Grundlagen der Geo-Informationssysteme. Wichmann Verlag, Offenbach. 866 pages.
- Grenzdörffer, G. (2003): Design and performance of the integrated digital remote sensing system PFIFF - experiences with urban applications. Proceedings of the ISRPS WG VII/4 Symposium, Remote Sensing of Urban Areas, Regensburg 27.-29.6.2003 (ISRPS Volume XXXIV – 7/ W9) pp. 66-71.
- Korduan, P. (2003): Standardisation in Data Management to Increase Interoperability of Spatial Precision Agriculture Data. Proceedings of the 4th European Conference on Precision Agriculture Berlin 15.-19.6.2003, Wageningen Academic Publishers. pp. 323-328.
- Piepel, C. (2002): Basiswissen Geodienste im Internet - Technologie und OGC Standards. in: 7. Münchener Fortbildungsseminar Geoinformationssysteme, 6.-8.3. 2002, Runder Tisch GIS e.V.
- Herrmann, C., Asche, H. (2001): Web Mapping. Wichmann Verlag, Heidelberg 2001.

<sup>53</sup> Yahoo Maps RSS

<sup>54</sup> <https://github.com/yohanboniface/GeoRSSToGeoJSON>

- Kropla, B. (2005): *Beginning MapServer. Open Source GIS Development.* Apress, Springer Verlag, New York.
- Mitchell, T. (2005): *Web Mapping Illustrated.* O'Reilly, Sebastopol.
- OGC (1999): *OpenGIS, Catalog Interface Implementation Specification (Version 1.0), OpenGIS Project Document 99-051.*
- OGC (2002a): *OpenGIS, Web Map Service Implementation Specification, Version 1.1.1, TC 211, Open GIS Consortium, Wayland, <http://www.opengis.org/techno/specs/01-068r3.pdf>.*
- OGC (2002b): *Styled Layer Descriptor Implementation Specification, Version 1.0.0, Editor: Lalonde W., Document: 02-070*
- OGC (2002c): *OpenGIS, The OpenGIS Abstract Specification. Topic 12: OpenGIS Service Architecture, Version 4.3, Open GIS Consortium, Wayland, <http://www.opengis.org/techno/specs/02-112.pdf>.*
- OGC (2005a): *OpenGIS, Web Feature Service Implementation Specification, Version 1.0, TC 211, Open GIS Consortium, Wayland, <http://www.opengis.org/techno/specs/02-058.pdf>.*
- OGC (2005b): *OpenGIS® Filter Encoding Implementation Specification. Version 1.1.0, Document: 04-095, [http://portal.opengeospatial.org/files/?artifact\\_id=8340](http://portal.opengeospatial.org/files/?artifact_id=8340).*
- OGC (2005c): *OpenGIS® Web Processing Service. Version 0.4.05-007 <http://www.opengeospatial.org/standards/requests/28>.*
- OGC (2006a): *Web Coverage Service (WCS) Implementation, Version 1.1.0, OpenGIS® Implementation Specification, Editors: Arliss Whiteside, John D. Evans, Open Geospatial Consortium, Document: 06-083r8, [https://portal.opengeospatial.org/files/?artifact\\_id=18153](https://portal.opengeospatial.org/files/?artifact_id=18153).*
- OGC (2006b): *OGC ® Sensor Web Enablement: Overview And High Level Architecture. Version 2.0, Editors: Mike Botts, George Percivall, Carl Reed, John Davidson, document: 06-050r2*
- Purvice, M., Sambells, J., Turner, C. (2006): *Beginning Google Maps Applications with PHP and Ajax. From Novice to Professional,* Apress, Springer Verlag New York.
- Ramsey, P. (2005): *The State of Open Source GIS.* Refraction Research Inc.
- Seip, C., Korduan, P., Zehner, M.L (2017): *Web-GIS. Grundlagen, Anwendungen und Implementierungsbeispiele,* Wichmann Verlag, Heidelberg, 552 Seiten.
- Strobl, J., Blaschke, T., Griesebner, G. (2002): *Tagungsband AGIT 2002,* Wichmann Verlag, Heidelberg.



### **Acknowledgements**

Our special thanks go to Dr. Alexander Born, Dr. Annette Hey, Dr. Meike Klettke, and Dr. Peter Korduan for their contributions to the first version of this textbook in 2007, which are used for an update and extension in this version.

This textbook also refers to our German textbooks on Surveying (Resnik/Bill, 2018), GIS (Bill, 2016) and Web-based GIS (Seip, Korduan, Zehner, 2018) and includes parts of our remote lectures in Remote Sensing (prepared by Dr. Görres Grenzdörffer) and Cartography (prepared by Dr. Annette Hey).

Rostock University  
Faculty of Agricultural and Environmental Sciences  
Chair of Geodesy and Geoinformatics  
Justus-von-Liebig-Weg 6  
D-18059 Rostock  
Germany  
Tel ++49-(0)381-4983201 (Secretary)  
Fax ++49-(0)381-4983202 (Secretary)  
igg@uni-rostock.de  
<http://www.auf.uni-rostock.de/gg>

