(Geo)database and Data Management

An introduction to the possibilities of a centralized (spatial) data storing and access as well as its data management constraints

(GLOWA Volta Project)

GLOWA Volta Project Phase III Inception Meeting and Workshop

Accra, Ghana, September 25th – 26th
Overview

1. Objectives within the database and data management project in phase III
2. Data within GVP
3. Data management
4. Database
5. Geodatabase
6. (Geo-)Database use scenarios
Objectives

- The main objective of the database project in Phase III is to build up a database which
  - reflects the research results on data level
  - is completed with regard to the data actually needed
  - is consistent, means that data quality is verified
  - is standardized, means research outputs are compliant with the requirements of following data users (with regard to e.g. measure units or formats)
  - is well described on metadata level, so that users can find data they need easily
  - has interfaces for data transmission (download, upload)
  - is based on technology, which is adjusted to the technical and scientific requirements of the transfer counterparts
Due the fact that the GVP is an interdisciplinary project, the data stock is very heterogeneous.

The datastock consists of structured (e.g. tables, ASCII files, geodata) and unstructured (documents) data.

Sources of data are:
- field investigations (digitized)
- measurements from monitoring stations
- research activities and own investigations
- institutions
What is the situation of the GVP Data?

- Data contents are

  - Socio-economic data about water use, landuse, migration, income sources, agricultural production (e.g. within common sampling frame)

  - Remote sensing data (rasters): e.g. clima, landcover

  - Geodata as vector or raster: administrative, infrastructural, natural resources: soil, geology, vegetation, hydrography, topography, landuse

  - Proprietary model data using the above mentioned data as input data and producing output data which are subjoint to the datastock
A great part of data is georeferenced by

- textual describing (research area is in town x)
- xy-coordinates in a table
- GIS-formats as raster or vector data

A wide range of formats used within the GVP is given (examples):

- Raster formats:
  - .tiff, .gif, .img, ESRI-grid, .rst, .jpeg, .bmp, .asc, .tin, .rds
- Vector formats
  - Shape, coverage (ESRI)
- Table or database formats
  - .exc, .sav, .txt, .mdb, .csv, .dbf, .dat
- Unstructured formats
  - .doc, .txt
1. Objectives
2. Data
3. Data management
4. Database
5. Geodatabase
6. Geodatabase use scenarios

Problem

- In its current state the GVP-datastock could be organized more adequately for an efficient using
- Yet, we don’t have a complete overview about the datastock of GVP
- Data is partly distributed and not centralized
The solution is: to improve our Data management

- Definition (by the „Data Management Association“):

  „Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise“
Components of data management

Procurement
- Own investigation
- Own processing
- Supply from other institution/Project

Structuring (Data modeling)
- Categorisation
- Sortation
- Description and Storing

Administration
- User Access Rights
- Security

Use and Processing
- Data processing
- Content Management
- Quality Assurance
- Data preparation (for others)

Live-Cycclus of data and his management

Disposal
- Update or
- Erasure

Distribution
- Access
- Deliver
Data management

• The (physical) core of data management is the database
• The database has to be embedded in a standardized data management environment
• This means
  - The database must have an implemented data model which fits the project requirements (see later)
  - Data using and flow has to be transparent with dedicated responsibilities regarding quality control, data structure standards and a consensus about data policy before data can be included in a database
• But what is a database?
Database: Definition

- A database is a collection of logically related data designed to meet the information needs of one or more users

Characteristics

- A database is managed by a database management system software (DBMS). A DBMS includes among others
  - tools for organizing the data inside the database (database structure/data model)
  - data query, retrieve and update features
  - features for maintaining security and integrity of data
  - backup, recovery and archiving tools
  - application programming interfaces allowing to connect other programs/models to the database
Remarks

• database is not inevitable equal to database
• There are a lots of different databases with different DBMS
• A database can be a CD as well as a powerful and difficult to administrate database system with Oracle as DBMS
Data Model

- Creating a conceptual data model is the first step in database design. It describes the objects (informations, data) that are represented within an organization and the relationships among them on an abstract level.

- A conceptual model has to be devolved into a logical model which can be implemented into a database.
Data Model

- A data model for the GVP which has to be developed, should at least meet the following criteria
  - storage of different data formats orientated at the models in use
    - E.g. differentiated in raw data (input data), output data and common exchange formats
  - storage of basic data for research issues
  - storage of research results in unstructured formats (documents)
  - storage of present geodata in a structure which enhances the information output of the present
Next steps within the database project

- Collection of the given GVP datasets which are missing
- Evaluation of the datasets (quality, further use)
- Development of standards for data use in terms of
  - methods: e.g. measure units
  - metadata: elements which have to be filled up
  - formats
  - responsibilities within the scientific disciplines
  - use permissions
- Development of a data model
- Analysis of data distribution (clients, presentation to the public)
- Decision making with regard to the technical demands
- Installation of the database and transfer of data into it
Geodatabase: Definitions

- A database is a collection of logically related data designed to meet the information needs of one or more users.

- A GIS database (Spatial Database or Geodatabase) includes data about the spatial locations and shapes of geographical features recorded as points, lines, areas, pixels, grid cells, or TINs, as well as their attributes.

- One of the main views of the geodatabase is as a data model for representing geographical information using standard relational database technology. A second, equally important aspect, is the way the geodatabase manages data access and storage.
Because ESRI-GIS is the standard software within the GVP it is recommended to use the Geodatabase from ESRI for storing geodata!!

**Why use the geodatabase from ESRI?**

- There are several geodatabases offered from other providers:
  - Oracle spatial
  - PostGIS (Open source/free software)
  - Geomedia SQL Server Warehouse Geodatabase
  - …

**Because ESRI-GIS is the standard software within the GVP it is recommended to use the Geodatabase from ESRI for storing geodata!!**
What can be stored in a geodatabase?

- A geodatabase has its own formats
- All geodata imported are converted into this formats
- ArcGIS Data Interoperability Extension allows import of more than 75 GIS-formats and export in 50 GIS-formats

- **(Non-spatial) Object class**: non georeferenced objects (table)

- **Feature Class**: georeferenced vector based objects, sharing the same geometry type, attributes and spatial reference. Corresponds with a shape-file

- **Feature Dataset**: Data Container containing feature classes. Feature classes have the same coordinate system and spatial extend. Comparable with an ArcInfo workspace
What can be stored in a geodatabase?

- **Raster Datasets**: standalone raster datasets or organized in Raster Catalogs
- **TIN Datasets**: triangulate irregular networks
- **Annotation Feature Class**: stores text or graphics providing information about features or general areas of a map
- **Dimension Feature Class**: a dimension feature class in ArcMap shows specific lengths or distances on a map
- **Geometric Network**: Edge and junction features that represent a linear network such as a utility or hydrologic system in which the connectivity of features is based on their geometric coincidence
- **Locators**: e.g. a table which contains x/y coordinates can be used for creating georeferenced point-features
Characteristics of the geodatabase

- A central storage of geodata with their annotations, frequently used methods (toolsets) and models

- The geodatabase data model brings a physical data model closer to its logical data model. It gives the objects behaviour by establishing rules like ....

  - that the values you assign to an attribute fall within a prescribed set of permissible values

Realized by definition of Attribute Domains
3. Geodatabase

Characteristics of the geodatabase

- …. establishing rules like
  - that a feature can be placed adjacent or connected to another feature only when certain constraints are met

- that a the geometry of features follows ist logical placement

Realized by definition of topology between geometric features
Characteristics of the geodatabase

- that collections of certain features conform to their natural spatial arrangements. A river always flows downstream

Realized by designing of geometric networks
Characteristics of the geodatabase

- Except for the Attribute Domains the other topological rules mentioned above are based on relationships among geometric objects.

- There are two main kinds of relationships in a geodatabase:
  - spatial relationships (associations) among geometric features
  - general relationships among geometric and non-geometric features (tables)

- E.g. a general relationship connects a table of census data with a feature class containing towns as point features.

- Relationships not only display dependencies but also have behaviour: e.g. when a borehole stops its water supply the connected irrigation system should be inactivated automatically.

ArcGIS provides tools for fixing topological errors within a geodatabase.
Summary of the benefits of the geodatabase data model

• A uniform repository of geographic data. All data can be stored and centrally managed in one database.

• Data entry and editing is more accurate. Fewer mistakes are made because most of them can be prevented by intelligent validation behaviour.

• Features have a richer context. Topological associations and general relationships define the context of features with other features and what happens when a related feature is moved, changed or deleted.

• Shapes of features are better defined by using straight lines, circular curves, elliptical curves …

• Sets of features are continuous. Geodatabases can accommodate very large sets of features without tiles or other spatial partitions.
Handling of raster data within a geodatabase

- A great part of the geographic data stock of GVP consists of raster data
- The possibilities of using raster data depend on the geodatabase type personal or enterprise geodatabase (explanation see below). Personal GDB don’t have full functionality as described below!
- Main advantages of working with rasters in a geodatabase
  - large data holdings can be easily built, modified, and utilized
  - tiling, indexing and pyramiding storing parameters allows to get only the tiles necessary to satisfy the extent and resolution of the query
  - compression makes it possible to store large, seamless raster datasets and raster catalogs (as large as several terabytes) and display it quickly to a client
Handling of raster data within a geodatabase

- Main advantages of working with rasters in a geodatabase
  - choice of creating mosaics or raster catalogs
  - time series can be showed using a geodatabase raster catalog
### Types of ESRI-Geodatabases

There are different versions of geodatabases from ESRI

**ArcGIS 9.1**

- **Personal Geodatabase**
  - based on MS Access
  - can be read simultaneously by several users
  - only one user at a time can edit the same data
  - no storing of features sorted in ongoing hierarchy comparable to a file system
  - 2 GB Limit

- **Usable for:**
  - personal use or in small workgroups (single user edit)
  - primarily storing of vector-data
  - managing raster data inside by storing outside of the geodatabase
Types of ESRI-Geodatabases

- **ArcGIS 9.1**
- **Enterprise ArcSDE Geodatabase**
  - based on Oracle, IBM DB2 and others
  - read/write by unlimited number of users
  - conflict-management by versioning
  - archiving
  - unlimited size

- Usable for:
  - use in great organizations with multiple users at the same time in a framework of long transactions
  - for large amount of raster data (as described above)
Access to the geodatabase

- The way of access to the geodatabase is related to
  - type of network (Local Area Network or Internet)
  - type of geodatabase

- and the type of the geodatabase is related to
  - the requirements of the GVP regarding to
    - their data related to the information level
    - the processes which will be effected in Phase III and after transferring data and methods to the stakeholders
Geodatabases in the ArcGIS Context

1. Objectives
2. Data
3. Data management
4. Database
5. **Geodatabase**
6. Geodatabase use scenarios

Data formats

- Coverage
- Shape
- Personal Geodatabase
- Server Geodatabase

GIS-Applications

- ArcView
- ArcEditor
- ArcInfo

Database Gateway

- Read only
- Read/Write

Internet-Browser

ArcSDE

ArcIMS

Map Server

Map Server

Database and data management project in GVP phase III
Access to the (geo)database: Scenarios

Scenario I: File Server

Conditions:
- Within a LAN (ZEF) and a File-Server as database
- File Server is part of an distributed resources environment and obtains only a part of the entire GVP data stock
- Data (spatial/non-spatial) is sorted in directories e.g. by themes
- Metadatabase is complete with accurate descriptions
- Clients: models running at ZEF, ArcGIS and other applications
Access to the (geo)database: Scenarios

Access to data:

- Geodata, model data and unstructured data (documents) can be searched by ArcGIS metadata functionality (using ArcCatalog) respectively by web based metadata service or File Explorer
- Copy of data from server and local storing for processing
- Upload of data based on standards regarding quality and control criteria for data controlled by an administrator from outside
- searching by web based metadata service
- remote access to database by data transfer internet protocol with given permissions by an administrator
Scenario I: File-Server

- **Data Server**
  - **Metadata Base**
  - **Data Flow**
  - **User Activity**

- **Local**
  - **Web Browser**
  - **GVP Member**
  - **ArcDesktop**

- **External User**
  - **Web Browser**
  - **GVP Member**
  - **WinSCP**

- **File Server**
  - **Data 1 (Project 1)**
    - Vector-File
    - Raster
  - **Data 2**
    - Vector
    - Raster

- **ZEF Area (LAN)**

**Database and data management project in GVP phase III**

**Model/Tools (Mike Basin)**

**Working Area**
Access to the geodatabase: Scenarios

Scenario I: Advantages

- Data storage on a central place
- Different formats can be stored without converting in another database-format
  - Relatively small administratory effort on the technical level
  - Low costs
  - The software which is used for the data use (e.g. only ArcView 3.x) plays only a minor role
  - Disc space to store data depends on the server capacity
Access to the geodatabase: Scenarios

**Scenario I: Disadvantages**

- The collective processing of data at the same time within one project is not possible
- No spatial defined queries to the database
- There are no integrated automatisms that ensure data integrity
- Data is connected without integration of the object behavior
- Spatial topologies are only based on and within ArcInfo-coverages
- Storage of huge seemless raster data sets without parametrized compression, tiling or pyramiding
**Access to the (geo)database: Scenarios**

**Scenario II: Multiuser (Geo)Database**

**Conditions:**

- within a LAN (ZEF) and a powerful Server (e.g. Oracle) as database
- Oracle Server is part of an distributed resources environment but optains the entire GVP data stock
- Implemented Geodatabase (as an own data model inside the database) is managed by application server ArcSDE (Spatial Database Engine)
- Structured data (spatial/non-spatial) is organized in a complex data-model and stored in the relational database format and managed by a database service
- Unstructured data (documents) is stored outside on a File Server
- Metadatabase is complete with accurate descriptions and connected to the main database
- Clients: models running at ZEF, ArcGIS as ArcSDE client and other applications (by programmed interfaces)
Access to the (geo)database: Scenarios

Scenario II: Multiuser (Geo)database

Access to data:

- Geodata can be searched by ArcGIS metadata functionality (using ArcCatalog) or by webbased metadata service
- Data of other formats can be searched by database queries or by web based metadata service
- Data editing inside the database (instance) directly by multiple users
- Conflict management by ArcSDE (for geodata) or the RDBMS and executed by an administrator
- Specific non spatial model data has to be im- and exported by a format converting process
- From outside by connecting the SDE application server
3. Data Management

Scenario II: Multiuser Database

- **Metadata**
  - base
  - GVP-member
- **data server**
- **data flow**
- **user activity**
- **working area**
- **GDB**
- **ArcSDE**
  - GVP-member/external user
- **Web-Browser**
  - direct connect
  - data download (mail)
  - edit/query
- **Arc-Desktop**
  - (ArcEditor)
  - local
- **query/search/edit**
- **ZEF-area (LAN)**
  - notice
    - contains:
      - attribute domains
      - spatial domains
      - topology
      - spatial datasets
      - validation rules
      - versioning
- **Record**
  - upload/download
  - notice contains:
    - feature dataset
    - object class
    - relation class
    - geom. network
    - raster dataset
- **Model**
  - GVP-member
  - record
  - notice contains:
    - attribute domains
    - spatial domains
    - topology
    - spatial datasets
    - validation rules
    - versioning
- **RDBMS (Oracle)**
- **External area**
- **Personal GDB**
- **Scenario II: Multiuser Database**
  - Personal GDB
Access to the (geo)database: Scenarios

Scenario II: Advantages

- Data storage at a central place and in a common format
- Multiuser-editing
- Support of long lasting transactions
- Topologys, geometric networks, object behavior
- Validation for maintaining data integrity
- Storage of huge amount of data, such as connected raster datasets (e.g. time series) within a database. Compression of rasters and high display performance
- Connection of fitting application to the database via ArcSDE

Disadvantages:

- Complicated installation and import of given data
- High administration effort and high costs for the RDBMS and its administrator as well as for the ArcSDE license
**Access to the (geo)database: Scenarios**

**Scenario III: Personal (Geo)Database combined with a File Server**

**Conditions:**

- within a LAN (ZEF) and a Personal Geodatabase for geodata and a fileserver for non ESRI-Formats as well as documents
- Rasters are mostly managed inside the geodatabase but stored on a file server
- Unstructured data (documents) is stored outside on a File Server
- Heterogeneous Data formats needed by models are stored on a File Server
- Metadatabase is complete with accurate descriptions and connected with the main database (geodatabase + FileServer)
- Clients: models running at ZEF, ArcGIS and other applications
Access to the (geo)database: Scenarios

**Scenario III:** Personal (Geo)Database combined with a File Server

**Access to data:**

- Geodata can be searched by ArcGIS metadata functionality (using ArcCatalog)
- Data of other formats can be searched by file explorer or by webbased metadata service
- Data for processing is downloaded and stored local. Geodatabase objects are stored as copyys in local Personal GDB (maintaining the geodatabase features like relationships, topologies)
- Upload of data based on standards regarding quality criteria for data and executed or controlled by an administrator
Database and data management project in GVP phase III

Scenario III: Personal GDB with File Server

- **Metadata base**
  - Contains:
    - attribute domains
    - a spatial domains
    - topology
    - spatial datasets
    - validation rules

- **Personal GDB**
  - Feature dataset 1
    - Feature class
    - Object class (Relation. class)
    - (Geom. network)

- **File System**
  - data 1 raster (remote sensing)
  - model data dokument
  - data 2 raster

- **WinSCP**
  - data

- **GVP-Interface**
  - data

- **ArcView 3.x**
  - external area
  - WinSCP
  - edit

- **ZEF area (LAN)**
  - local

- **Web-browser**
  - search
  - update

- **ArcDesktop (ArcEditor)**
  - local

- **GVP member**
  - search
  - update

- **model/tools (Mike Basin)**
  - search
  - update

- **data server**
  - record upload/download
  - download per mail

- **data flow**
  - user activity

- **working area**
Scenario III: Advantages

- Data storage at a central place and in a common format
  - Low costs
  - Simple administration, especially with ArcInfo clients with an own PGDB
  - Topologies, Geometric networks, relationships
  - Validation for maintaining data integrity
  - Simple labeling through annotation-classes
  - Dimensioning
  - Model formats can be hold in his own formats
Scenario III: Disadvantages

- The collective processing of data at the same time within one project is not possible
- Storage space of the PGDB is limited. Especially raster data can be integrated to a limited extend only
- No storage of huge raster data sets and no download of selected and compressed tiles from raster data from storing place
- No support for long lasting transactions (disconnected processing of parts of the database and controlled merging by the GDB)
**Access to the (geo)database: Scenarios**

**Scenario IV: Personal (Geo)Database combined with a File Server and a Map Server**

**Conditions:**
- As Scenario III ..... but additional
- Map server ArcIMS linked with a web server providing geodata and maps
- Map services offered within the metadatabase or the GVP homepage

**Access to data:**
- Within the LAN as Scenario III
- Map services of different kinds offered within the metadatabase or the GVP Homepage
- Searching and geodata extraction from the geodatabase using map services (extract service)
Database and data management project in GVP phase III

ArcIMS + Webserver

Maps

external institution/project: biota, DLR

data

ArcDesktop
(ArcEditor)

ArcReader

ArcExplorer

HTML/Java-viewer

GVP-Homepage

Feature Service

ArcMap Image-Service

catalog Service

search

Image Service

data

local data

data extraction

data flow

user activity

working area

ZEF: server area

file system

shape

coverage

raster

documents

maps

ArcIMS + Webserver

Personal GDB

Feature dataset 1

Feature class

Object class

Relation. class

Geom. network

Raster dataset

author tools (ArcMap-Server)

GVP member

stakeholder

external User

 Hijack 1 view

view

stakeholder

External User

html/Java-viewer

ArcMap Image-Service

ArcDesktop (ArcEditor)

data upload (mail, ftp ...

data flow

user activity

working area

ZEF: local area

arcspace  

shape

coverage

raster

documents

author tools (ArcMap-Server)
Scenario IV: Advantages:

- GVP results can be presented in a cartographical demonstrative way to a broad audience.
- Data users get an overview over existing maps and data and can order or download diverse geodata.
- Maps and data from other producers, who also offer internet map services can be integrated in ArcDesktop products and thus be combined and processed together with data of the GVP project and/or local data.
- A metadata catalog can be implemented besides the existing metadatabase.