INNOVATION OF THE ARCČR 500 DATA MODEL AND ITS CONVERSION INTO ESRI MAP TEMPLATES

Petr Čejka[1], Pavel Hájek[2], Karel Jedlička[3]

[1,2,3] Section of Geomatics, Department of Mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 22, 306 14, Plzeň, Česká republika
petr.cejka6@gmail.com, pavel.hajek21@gmail.com, smrcek@kma.zcu.cz

Abstract

A need of conversion of existing data into new structures and for other purposes arises in the age of geographic information systems and large geographic databases produced by various organizations and companies. There is also often a need of harmonization of two discrepant data sources into a third one. The paper is thus focused on description of tools suitable for above-mentioned activities, particularly: Free Mind, Geodatabase Diagrammer, ArcGIS Diagrammer, ModelBuilder and ArcGIS Desktop.

These tools are described on two projects realized by authors of the article: “Innovation of the ArcČR 500 data model” and “Conversion of the innovated ArcČR 500 data model into the structure of ESRI Map Templates”. The aim of the innovation of the structure of the existing ArcČR 500 geographic database was the conversion of its flat structure into appropriate vertical and richer horizontal structure. The primary data base was created as a flat file structure and the first really geodatabase version was the version 2.0, where datasets and relationships were used. However, the created version 2.1 has a re-structuralized vertical structure and consists of data structures whose allow better exploitation of both spatial and attribute part of the geographic database, namely: more relationships, created topologies, geometric networks, terrain and raster datasets, etc.

The second, following project was focused on conversion of created geographic database ArcČR 500 v 2.1 to several cartographic databases of different scales, whose can be cartographically visualized according to ESRI Map Templates. The aim of this project was to provide the data from Czech Republic into small and medium scales of the ESRI Map Templates.

ArcČR 500 data model vision 2.1 has been currently passed to ARCDATA PRAHA Company, which considers the use of the created data structure for a new version of their ArcČR 500 database. This ArcČR 500 version is planned to publish after data actualization. The data import from ArcČR 500 to ESRI Map Templates is currently (2010 fall) consulted with ESRI.

The goal of the contribution is to present the potential of existing data modeling tools by demonstration of realized projects. The presentation shows, that the core of the data modeling is not in the development of newer and better tools, but in the design of high quality conversion/harmonization model (by using already existing tools).

Authors cooperated with the Czech ESRI distributor (ARCDATA PRAHA) on both projects. The third author was supported by the Research Plan MSM 4977751301. Authors have also used their experiences from their work on data models for themes from INSPIRE annexes, which have gathered from Plan4all project (ECP 318007).

Keywords: ***data modeling, harmonization, INSPIRE, Free Mind, Geodatabase Diagrammer, ArcGIS Diagrammer, ModelBuilder, ArcGIS Desktop.***

THE CURRENT SITUATION OF DATA MODELING IN EUROPE

On the present, the INSPIRE directive (more on JRC (2010)) is a leading initiative, which influences architectures of both existing and newly built spatial databases in European Union. INSPIRE directive addresses 34 spatial data themes, whose are described in three Annexes. There are already existing guidelines and data models for themes addressed in Annex I and there are Inspire Thematic Working Groups working on guidelines and data models for themes described in Annexes II and III. Actual results of continuous development in described area can be found at web pages:

- Guidelines: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2
- UML Model: http://inspire-twg.jrc.ec.europa.eu/inspire-model/

The current situation in European Union is that most of spatial database administrators are, or are going to be, forced to make their geographic data structures interoperable to INSPIRE data models. There can be seen two or three major scenarios.

The first scenario consists in one time data conversion from existing data model to an INSPIRE compliant data model. The second scenario lies in repeated conversion from source to target data model and the conversion can be done in a batch mode or in real time ("on the fly"). Of course there could be also third scenario, which means to built a database from scratch, even if it is rare to build a whole database on a Greenfield.

It is important to deeply discover the data structure of both source and target data models, to understand metadata and the data semantics. These are crucial points whose have to be deal right to successfully undertake the data design or data harmonization. Tools which are used can vary and they are works as the technical background.

Following chapters present two projects recently realized by authors of the article: "Innovation of the ArcČR 500 data model" and "Conversion of the innovated ArcČR 500 data model into the structure of ESRI Map Templates". The aim of the presentation of these projects is to show them as possible best practices of data model creation or conversion, pointing out the important steps of such a process.

INNOVATION OF THE ARCČR 500 DATA MODEL

The geodatabase ArcČR 500 is a spatial or geographic database in the level of detail 1 : 500 000. Its first version arises out of the classic paper maps of the Czech Republic, whose were digitized into a vector format. Until the version 1.3, the database had just a flat file structure. The actual distribution version 2.0 is stored in ESRI Geodatabase format and takes advantage of a classic Entity-Relationship-Attribute model (ERA model). More about the history of ArcČR can be found in Čejka (2010) or ARCDATA (2010). The ESRI Geodatabase structure is deeply described e.g. in Arctur & Zieler (2004).

The goal of the project *Innovation of the ArcČR 500 data model* was to convert the existing database structure into a new one, which would take the advantage not only from ERA model, but either of other possibilities of the ESRI Geodatabase data format, whose are summarized in figure 1.

At the start, the ArcČR 2.0 had still flat structure, which was just converted to a database environment. There was just a division into three datasets: admin_clen (administrative division), klady (map layouts) and mapove_prvky (map elements) and there were relationships used in the dataset of administrative division, see left part of fig. 2.

During the conversion process, the approach of a database development described in Longley et al. (2005) was followed:

1. Conceptual Model: model the user view, define objects and their relationships and select geographic representation.

2. Logical Model: match to geographic database types and organize geographic database structure.
3. Physical model: define database schema.

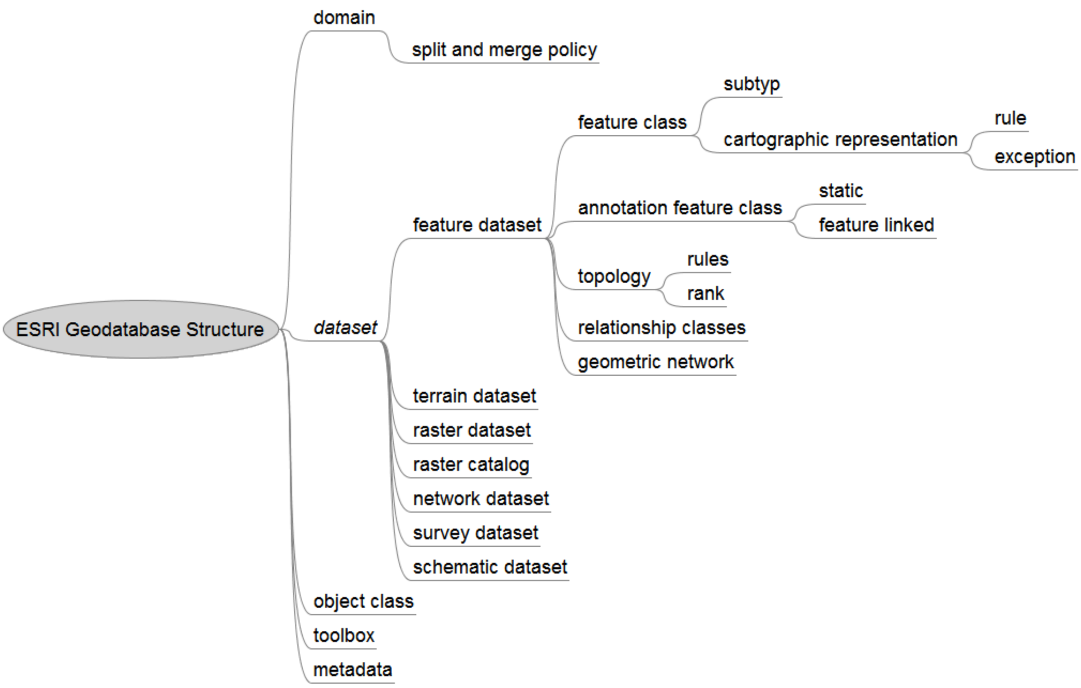


Fig. 1. The structure of ESRI Geodatabase, adopted from Jedlička (2010).

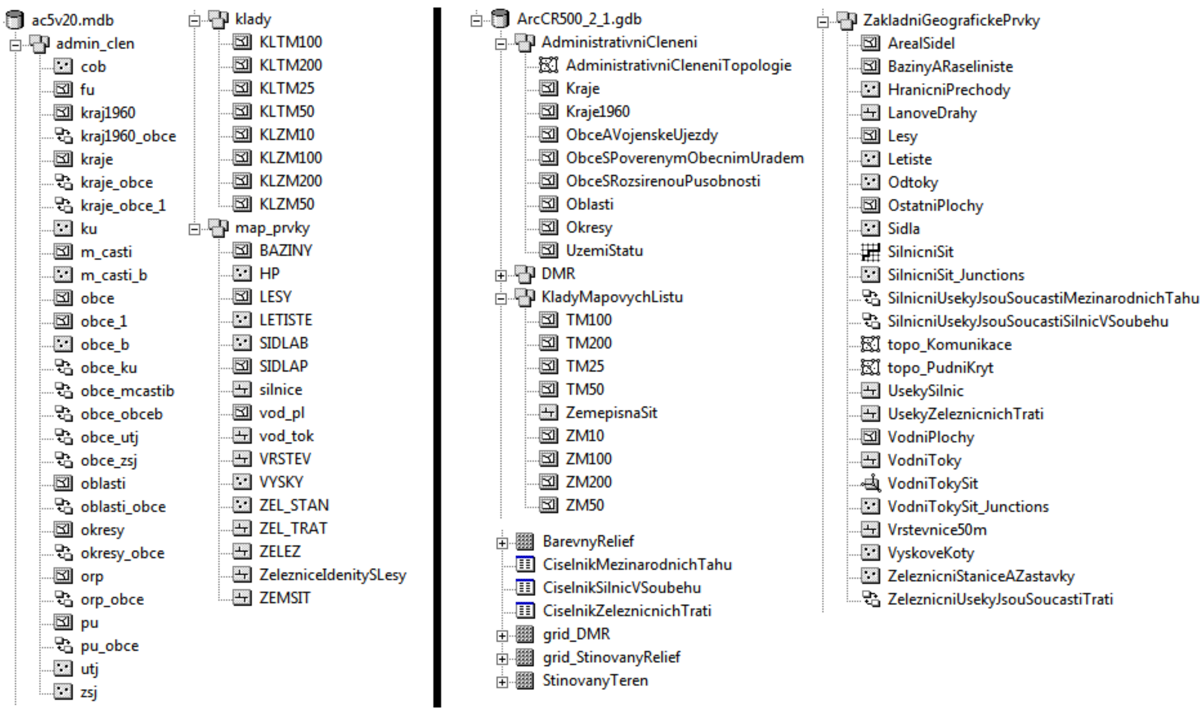


Fig. 2. The starting (at left) and resultant structure of ArcČR 500 (at right).

Naturally some steps were skipped, because in this case it was not about creating a database from scratch. There were some existing data before, even if they had just a flat structure:

The User view: the ArcČR 2.0 was primarily developed and abundantly used for creating of thematic cartographic outputs. Even the data structure allowed basic geographic analysis, the data structure was not sufficient e.g. for downstream analysis. Thus the purpose of development a new version of ArcČR was to improve its structure for geographic analysis, such as: hydrologic analyses, traffic analysis, DEM analysis, etc. Note: according to the database level of detail (LOD), analyses processed on the data have more educational than factual purpose.

At this step, the only used tools were brain, paper and pencil.

Objects were already defined and divided into classes, so their **geographic representation** and later their **geographic database types** were already set and were not changed (up to some exceptions in using specific data structures for both water and transportation networks or terrain).

The main deal was to **organize geographic database structure**, whose would be more convenient to analysis purposes. The whole process is deeply described in Čejka (2010), but the key steps follow.

1. First it was necessary to make a *deep revision of existing data and their structure*. It was necessary not only to understand the basic tree structure depicted at the left side of the figure 2, but also to understand to existing relationships in administrative division dataset (there was figured out that they represents spatial relations which can be in GIS modeled in a more suitable way – topologies).

   Next the attribute statistics had to be done and it was figured out that there was a lot of empty attributes (attributes were not added, because of lack of time, but there was proposed a method how to do such an attribute statistics based on Structured Query Language and mainly the SELECT DISTINCT construct.

   Also it is important to check if the data has some naming convention and choose one in a case of need. The UpperCasseForFirstLetters policy was selected and it was decided to use more descriptive names for datasets, feature classes and attributes. More about naming conventions in geographic database e.g. in Jedlička (2005a,b).

   The ArcGIS software and SQL were used for this step.

2. Second step consisted of a *proposal of new database structure* (the right part of fig 2 and appendix A). Two well known concepts related either to lexical or geographic databases were followed:

   - Database normalization – up to third normal form: Table faithfully represents a relation and has no repeating groups. No non-prime attribute in the table is functionally dependent on a proper subset of a candidate key. Every non-prime attribute is non-transitively dependent on every candidate key in the table (http://en.wikipedia.org/wiki/Database_normalization).
   - Topology prior to relations – spatial relations should be better represent by topologic rules than entity relationships, except the relationship is useful in often realized attribute queries. Otherwise the entity relationship can be anytime created from spatial relations among objects. Following this concept can save approximately one third of entity relationships in resultant data model (according to author's experiences).

   Thus there was proposed to delete relationships in the Administrative division dataset (newly named AdministrativniCleneni) and instead of them it was proposed to use a set of topologic rules. There was also proposed to delete redundant point classes whose represented reference points to polygons (relict from CAD boundary representation). Their attributes were proposed to transfer to appropriate polygon classes.

   Layers in the dataset of map layouts (KladyMapovychListu) stayed unchanged; the dataset was just supplemented by lines of geographic network.

   Biggest changes were proposed in the dataset of map elements which was renamed into BasicGeographicElements (ZakladniGeografickePrvky). As was already mentioned, al lot of feature classes was proposed to rename to more descriptive names. Next proposals:

- The geometry of roads and railroads should be strictly divided into segments from intersection to intersection. All following structures should be modeled using a many to many relationship: one segment can be a part of multiple roads and of course a road is composed of more than one segment. Additionally, the network dataset structure, which allows an analysis of transportation networks, was proposed to use for road feature class (SilnicniSit).
- Each river has to have a centerline representation. Each centerline has to be downstream oriented. The geometric network structure should be used (for purposes of hydrologic analysis).
- A land cover of the Czech Republic should be completed. Thus it is necessary to add the feature class Other Areas to already existing classes UrbanizedAreas, Woods, WaterAreas and Swamps,
- All above mentioned feature classes has to be of course topologically cleared and with set up topologic rules for further maintenance

Last but not least, the ESRI geodatabase format allows storing raster (raster dataset and raster catalog) and hierarchic triangular irregular networks (terrain dataset). Thus the digital terrain model and imagery should be uploaded directly to database.

The logical model of a geographic database has to be a result of this step. Although it could be still just drawn on a paper, the better way is to use some modeling tool. There of course exist tools based on unified modeling language (UML), but it is not a focus of this paper. Authors used the Geodatabase Diagrammer tool (http://arcscripts.esri.com/details.asp?dbid=13616), which creates an ERA model (including topologies) from existing geodatabase and stores it in Microsoft Visio. The Visio environment allows editing the model and thus proposing the new model.

Unfortunately there does not exist any way to convert the proposed model back to a geodatabase format. This could be seen as a disadvantage, but in many cases is the logical mode still just too much simplified to be used as a blueprint of a database schema. Nevertheless, a two-way tool: ArcGIS Diagrammer (http://kiwigis.blogspot.com/search/label/ArcGIS Diagrammer) can be used.

The last part of the project was to **define the database schema**. It consisted of creating of physical structures of each particular proposal described in previous chapters. Its description would exceed the paper scope, but it is also described in Čejka (2010). Authors use a combination of interactive and batch ArcGIS tools for creating the resultant database schema of ArcČR 500 v 2.1. Mainly ArcCatalog was used for database structure changes, Python scripting for batch processes and also ArcMap for validating the results.

CONVERSION OF ARCČR 500 INTO ESRI MAP TEMPLATES

The following project consists of conversion data between two databases, both with well described structure. The source one is the above described ArcČR 500 v. 2.1 and the target database is a set of geodatabases which underlies the ESRI Map Templates (http://resources.esri.com/maptemplates/).

ESRI Map Templates allows user to take advantage of predefined map design, but the user has to fulfill the condition, that the data has to be in prescribed data structure. In a case of ArcČR 500 v. 2.1 the ESRI Topographic Map Templates for small and medium scales were selected. The right part of a figure 2 shows the ArcČR 2.1 structure and the figure 3 depicts the structure of target geodatabase for small scales of topographic maps. Note that the target database also consists of conversion models for each its particular feature class. The visualization of the target geodatabase, using symbology of level of detail 9 (for scales closed to 1 : 1 000 000), is shown in appendix B.

While the left part of figure 3 depicts the prescribed data structure, the right part shows models whose were developed during the project. These models take the appropriate data from a source database and convert (e.g. combine, filter, transform, etc.) them into a required target feature class. The range of the contribution does not allow describing all of the models. Thus let us focus of one example on which it is possible to show the principles of conversion process.

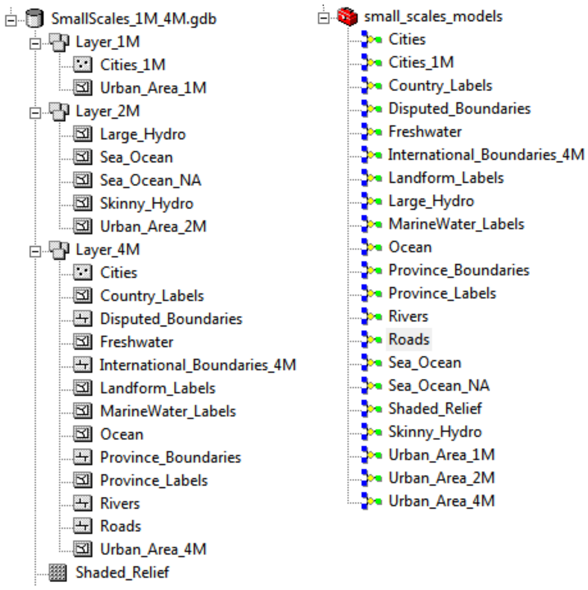


Fig. 3. Structure of geodatabase structure underlying to ESRI Topographic Map Template – Small Scales.

A road conversion model depicted in figure 4 is a moderately difficult case. Before the model conversion begins it is important to understand the possible different data semantic between the source and target data model. The semantic problems in our project come mainly from differences between European/national and US data classifications. In this road example it means to match the different classification of roads. Czech terms highway(or motorway) ~ dálnice, speedway ~ rychlostní komunikace, primary roads ~ silnice první třídy, secondary roads ~ silnice druhé třídy, third class roads ~ silnice třetí třídy and other communications ~ ostatní komunikace had to be mapped to codes used in result database whose come from the US classification Motorways (interstate, freeways ~ level 0), trunk (arterial highways ~ level 1), primary (level 2), secondary tertiary, residential, unclassified and service roads. The mapping function for small scales is portrayed in the table 1. Even if the resultant conversion table is simple, the example shows the need of understanding the data semantic. Note the European road classification can be seen e.g. there: http://en.wikipedia.org/wiki/Hierarchy_of_roads and the American classification there: http://wiki.openstreetmap.org/wiki/United_States_Road_Classification.

Table 1. The mapping function for roads in small scales.

Road classification	ArcČR 500 v.2.1	Map Template – small scale
	Dálnice	Level 0
	Rychlostní komunikace	Level 1
	Silnice 1. třídy	Level 2

Processing of the model in figure 4 starts at its bottom left corner (and later goes from left to right), where a new feature class schema (Roads) is created from a template and stored in appropriate dataset (Layer_4M). Then the features matching level 0-2 are selected from source dataset (highways, speedways and primary roads) and consequently this division is mapped to corresponding levels. Finally other attributes are set to a right parameter (Nation=420 ~ the code of the Czech Republic and the Continent="EU" ~ stands for Europe).

The rest of feature classes (and one raster dataset of course) were built in the same way, which was demonstrated above and which can be divided into several steps:

1. First, it is necessary to understand the target feature class structure and semantic and then seek appropriate data in a source database. Of course, that there can happen a situation, that there is no match in source data.

2. Then a ModelBuilder environment is used for creating a model consisting of sequences of tools, which convert the data into desired structure. Although it is possible to create own script (in Python) or even develop own tool, in our project the standard tools from ArcToolbox were sufficient.
3. There could be a need of geometric match on boundaries of particular national data, while development of a seamless data structure. In that case some method of geometry alignment has to be done. There are generally two possible methods: interactive editing or non residual transformation. Because it was not a goal of this project, it is not described any further.

The used technology (ModelBuilder in ArcGIS) has the advantage that it can be easily implemented in a server environment. Therefore it is possible to offer it even as a kind of a web service, so the conversion can be done on demand.

The conversion process of the Map Template of medium scale is analogous, however models become longer and more complicated, because of a richer content of medium scale map.

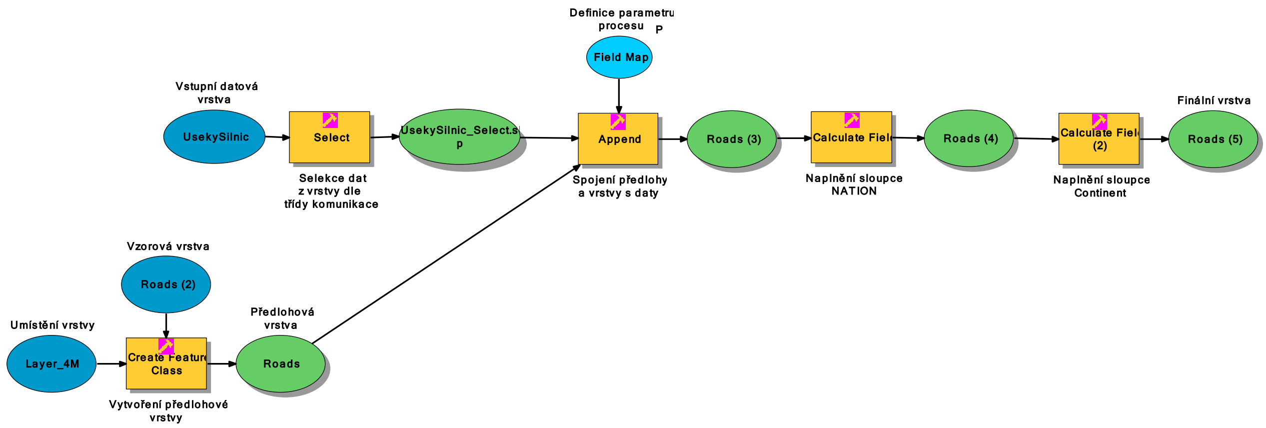


Fig. 4. Conversion of roads – an example of a ModelBuilder Model.

RESULTS AND SUMMARY

Results of project described in this paper are following: ArcČR v 2.1 was developed and well documented in Čejka 2010. ESRI Map Template for small scales was filled with data from the Czech Republic and medium scales have been analyzed, but no yet filled with data, see more in Hájek 2010.

First project shows a scenario of one time conversion process – creating a new structure of existing database (ArcČR 500). Second project show a conversion (harmonization if you want) mechanism from source to target database structure. The ArcČR 500 v 2.1 project has proved the conceptual~logical~physical stages of (geographic) database design. During the work on ESRI Map Templates project, the importance of data semantic and geometric matching appeared. Both of these projects (and several others, whose authors were cooperate, e.g.: data model for ZČE, City of Encinitas data model, geomorphologic database) show that existing tools are sufficient for data modeling, conversion and harmonization and that the core of this domains lies in deep understanding of the data and phenomena whose represent.

REFERENCES

Journal article:

JEDLIČKA, K. (2005a). Konvence v pojmenovávání geodatabáze. 1. část. In ArcRevue. roč.14. č.1. s.21-22. ISSN 1211-2135.

JEDLIČKA, K. (2005b). Konvence v pojmenovávání geodatabáze. 2. část. In ArcRevue. roč.14. č.2. s.25-26. ISSN 1211-2135.

Book:

Arctur, D., Zeiler, M. (2004). *Designing Geodatabases – Case studies in GIS Data Modeling*. Redlands. California. ESRI Press. ISBN 158948021X.

Čejka, P. (2010). *Inovace datového modelu ArcČR 500.* Západočeská univerzita v Plzni, Plzeň.

Jedlička, K. (2010). *Geomorfologický informační system.* Vysoká škola báňská, Technická univerzita Ostrava, Ostrava.

Longley, P., A., Goodchild, M., F., Maguire, D., J., Rhind, D., W. (2005). *Geographic information systems and science*. Paul A. Longley [et al.]. Chichester. John Wiley & Sons.

Technical documents:

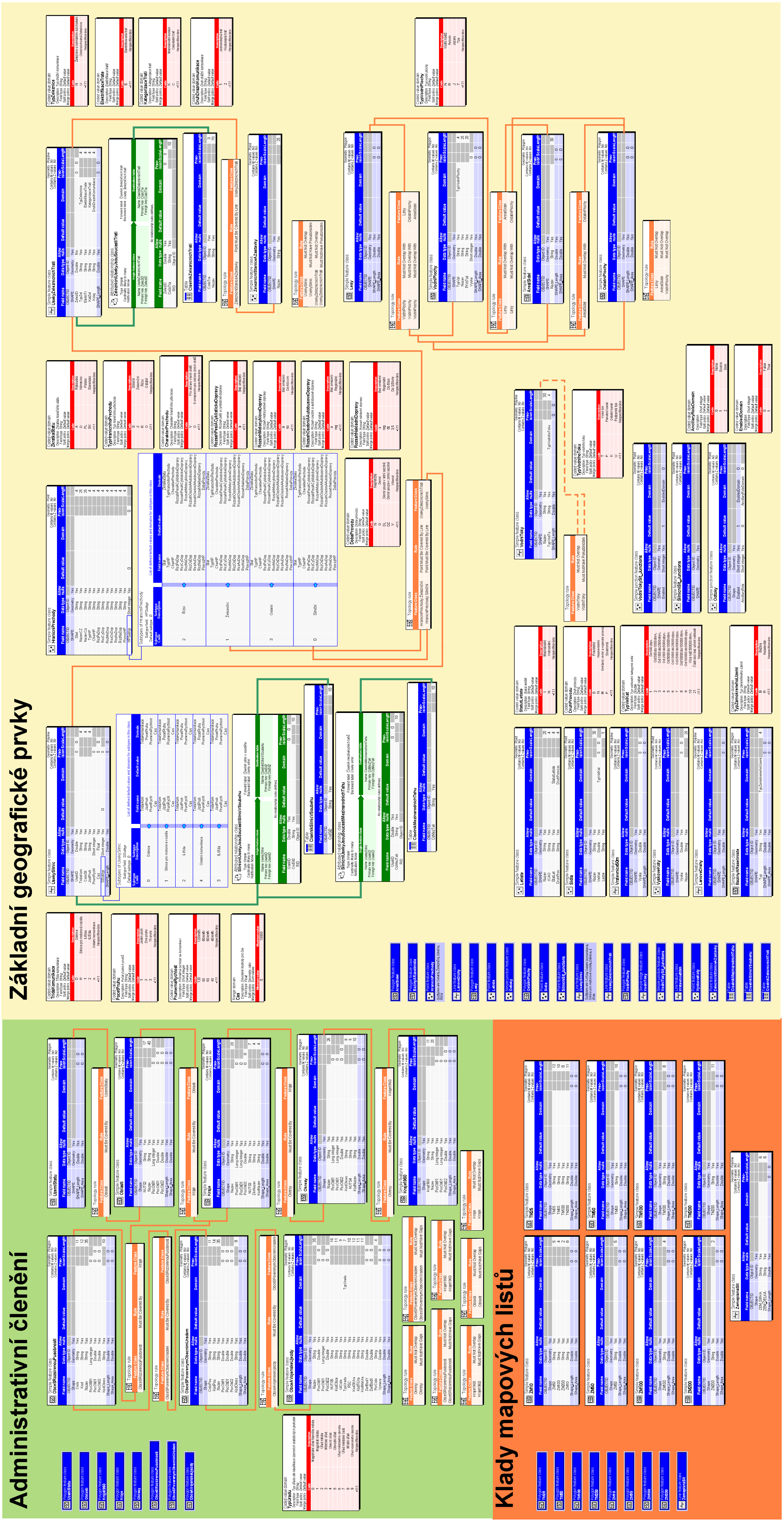
Hájek, P. (2010). Tvorba Map Templates pro databázi ArcČR 500 v. 2.1.Západočeská univerzita v Plzni, Plzeň,

Reference to a web page:

ARCDATA PRAHA, s.r.o., ArcČR 500. http://old.arcdata.cz/data/arccr. 11/6/2010.

European Commission Joint Research Centre (JRC). INSPIRE > Welcome to INSPIRE http://inspire.jrc.ec.europa.eu/. 10/25/2010.

APPENDIX A – THE STRUCTURE OF ARCČR 500 V. 2.1



APPENDIX B – TOPOGRAPHIC MAP TEMPLATE – SMALL SCALES