Characteristics in Data Management within a Scientific Multinational Internet Atlas

Christian RESCH

University of Vienna, Department of Geography and Regional Research
Universitätsstraße 7, A-1010 Wien, Austria
e-mail: resch@atlas.gis.univie.ac.at

Peter JORDAN

Austrian Institute of East and Southeast European Studies
Josefsplatz 6, A-1010 Wien, Austria
e-mail: peter.jordan@osi.ac.at

Abstract

The Atlas of Eastern and Southeastern Europe is a scientific atlas, which has been published in printed form since 1989. From April 2000 to July 2001 a prototype of an Internet version of this atlas was developed as a cooperation project between the institutes mentioned above. Compared to the many digital national atlases published recently, an Internet atlas covering various countries poses an unsuspected challenge in cartography within Internet communication. To accomplish this challenge it is necessary to consider the following specifications, which are discussed in this paper:

- ***Handling of complex data bases of multinational atlases****: It is necessary to access very different data sources. Therefore the first chapter of this paper is dedicated to the tasks of homogeneous data editing, data harmonization of inhomogeneous sources, consideration of national differences in classification, monitoring methods, monitoring dates, monitoring periods, reference areas etc.*
- ***Flexible conception of the Internet Atlas of Eastern and Southeastern Europe****: The open conception of the atlas follows a compromise between a restrictive and a flexible conceptual approach. The selection of the topics, map extents, etc. is not statically defined. New contents are produced as response to newly arising political, economic, social and environmental issues. Therefore it is necessary to offer an atlas structure with adaptable and manageable revisions and updates.*
- ***Preparation of the Internet atlas and its contents for an international auditory****: With a multinational Internet atlas conceived for an international auditory it is necessary to offer a multilingual user environment and to take special care of the handling of geographical names. Therefore the third chapter of this paper contains a disquisition about geographical names and the problems with special characters and diacritics in the Internet environment.*

Introduction

The Atlas of Eastern and Southeastern Europe (AOS) is a scientific atlas of a larger region. Since 1989 the Austrian Institute of East and Southeast European Studies publishes up-to-date maps in printed form as a map series. The atlas has the intention to offer up-to-date information about this part of Europe which was formerly Communist and is now in transition towards market economy and Western democracy integrating itself at the same time into European political structures. The information is presented in the form of thematic maps from the fields of ecology, economy and population. The main purpose of the atlas is to offer cross-regional comparisons by frame maps showing the complete region or larger parts of it in scales between 1:6,000,000 and 1:3,000,000. In this regard the concept differs from national atlases. However, the atlas contains also maps of individual countries and even parts of countries in larger scales, to focus more specific topics and case studies.

From April 2000 to July 2001 a prototype of an Internet version of this atlas is being developed under the project title AtOS as a cooperation project of the Austrian Institute of East and Southeast European Studies and the Department of Geography and Regional Research of the University of Vienna. The motivation for developing an Internet version of the atlas is to offer additional possibilities for the user, to publish information more up-to-date and to expand the auditory. Internet users who are interested in the post-Communist countries should be able to use the atlas as an information pool to get an overview of the conditions and relations in the region and as a gateway with links to more detailed and specialized information.

During the last years many digital national atlases were developed and some were also published via Internet. Technical possibilities are developed quickly and an ever larger number of Internet users utilize these cartographic information sources [Peterson (1999)]. In this paper, by the example of the Atlas of Eastern and Southeastern Europe, specifics will be discussed which should be considered with developing an Internet atlas covering various countries. Firstly, we will have a look at the handling of the complex data bases of this multinational region. Secondly, we will focus on the issue that the conception of the atlas must be in accord with a flexible data management. The fact that the preparation of an Internet atlas and its contents for an international auditory has a strong impact on data management will form the third and final part of this paper. In this context it is necessary to throw a special glance at the possibilities and limitations of integrating the special letters and diacritics of a wider range of Roman alphabets as well as of transliterations from non-Roman alphabets into an Internet environment.

1. Handling of the complex data bases of multinational atlases

In contrast to national atlases, regional atlases offer an overview over a variety of countries. This fact gives them a certain significance in a period of globalization and European integration, when international overview is urgently needed, particularly in the fields where integration and international cooperation have advanced or would be of ultimate importance (like environmental issues, economy and population development).

For producing a map covering a range of different countries it is necessary to take into account very different data sources. Even adjoining countries may record and document data with different methodological frameworks. They may ask for different sets of items, may apply different classification methods, methods of measurement, refer the data to different points in time etc. Strictly speaking, among the topics usually presented in national and regional atlases only a few can be documented by completely comparable data for a wider range of countries and may then be represented properly also in multinational atlases. National differences in the methods of data collection are, of course, not only a matter of insufficient international coordination and cooperation, but based on different cultures, living conditions, economic systems etc. and are thus in a way unavoidable. Therefore it is very often impossible and also worthless to show exact values on multinational maps. It becomes necessary to turn to qualitative classifications (for example "strong – medium – little") instead of quantitative indications [see also Jordan (1994), p.97ff].

With harmonizing inhomogeneous data sources it is important to consider the following possible differences:

- **Definition of terms, classification methods, thresholds, methods of measurement:** On the one hand there are little problems to expect with data of fields where international standardization has proceeded to a considerable extent (especially in natural sciences). On the other hand comparisons may be very difficult or impossible at non-standardized data (especially in human sciences).
- **Reference areas:** Data referred to monitoring networks of different density or to divergent sizes and structures of reference areas feature different degrees of accuracy and are not strictly comparable. The spatial dimensions, e.g., of political-administrative units of individual countries (regions, districts and communes) tend to be very different as well as their internal structures. In some countries, cities and rural areas are merged into one administrative unit, in others cities/towns form separate administrative units. Thematic data related to such different reference areas result in cartographic representations which are not really comparable.

- **Geometrical accuracy:** Sometimes data sources with very different projections, scales and geometrical generalization have to be used. It might be necessary to adjust generalization of some sources to homogenize cartographic representation.
- **Time reference:** With some topics like population density a common reference period of a couple of years may be sufficient. However, with topics heavily affected by sudden political, social or natural events, like transportation, migration or even tourism, a common point of time reference is indispensable for creating comparable maps.

To be able to get aware of differences in data quality and to harmonize the data it is important to have data collections containing a minimum of metadata (including methodological explanations). Unfortunately, this is very often not the case. The gap is most conveniently bridged by involving national experts as advisors or map authors. They will contribute to:

- Finding out differences in data quality. Especially necessary, if metadata and methodical explanations are insufficient.
- Evaluation and comprehension of the differences between data bases of different countries.
- Homogenization of the data, e.g. possible by transforming a quantitative into a qualitative classification.
- Scientific verification of the contents. This item is especially important because it is risky for an atlas editor to assume responsibility for the contents without verification by experts of the countries covered. It could be embarrassing if maps show only stereotypes.

Communication and cooperation with national experts is substantially facilitated by an Internet version of the atlas. The preliminary version of a map can be presented to the authors by non-public parts of the atlas. The authors are then able to harmonize and correct the data they are responsible for.

If it is impossible to harmonize data from different sources, the map editor has three possibilities [see also Jordan (1994), p.97ff]:

- If (comparable) data for larger central parts of the map section are lacking it is better to refrain from publishing the map.
- Marginal parts of the map section or smaller areas without (or with non-comparable) data could be marked by a special symbol ("no data").
- Within a certain range of comparability variations in data quality can be indicated by variation of symbols on the map or by indication on an inset map.

This is just to hint at the fact that although it has technically become quite easy to generate maps from digital data sources, data quality and data homogeneity within a map are still a problem and must be carefully observed, especially when representing multinational regions.

2. Flexible conception of the Internet Atlas of Eastern and Southeastern Europe

Most national and regional atlases are elaborated according to a predefined editorial concept that covers contents, scales and aerial sections of all maps [see also Neumann (1999)]. With this multinational atlas, however, the concept is open and flexible. A master concept defines just the main theme fields and the limits of the region to be covered. Contents, scales and aerial sections of individual maps depend on current interest arising from political, social, economic or environmental events and developments as well as on data availability [see also Jordan (1993)] (see also chapter 1).

An Internet atlas corresponds very well to this flexible and open editorial concept. It is able to react quickly and flexibly to current demands. Data editing is, of course, still a problem to be managed (see chapter 1), but then a map can be made accessible to the public without any restrictions, e.g. independent of delivery dates.

Notwithstanding these advantages as regards flexibility and publication, what is made accessible to the public is not be an unstructured accumulation of different maps. The concept is not as static as it was with printed atlases, but there still exists a concept. Theoretically, there are three possible approaches towards an open and expandable concept. Their basic structure is shown in Figure 1.

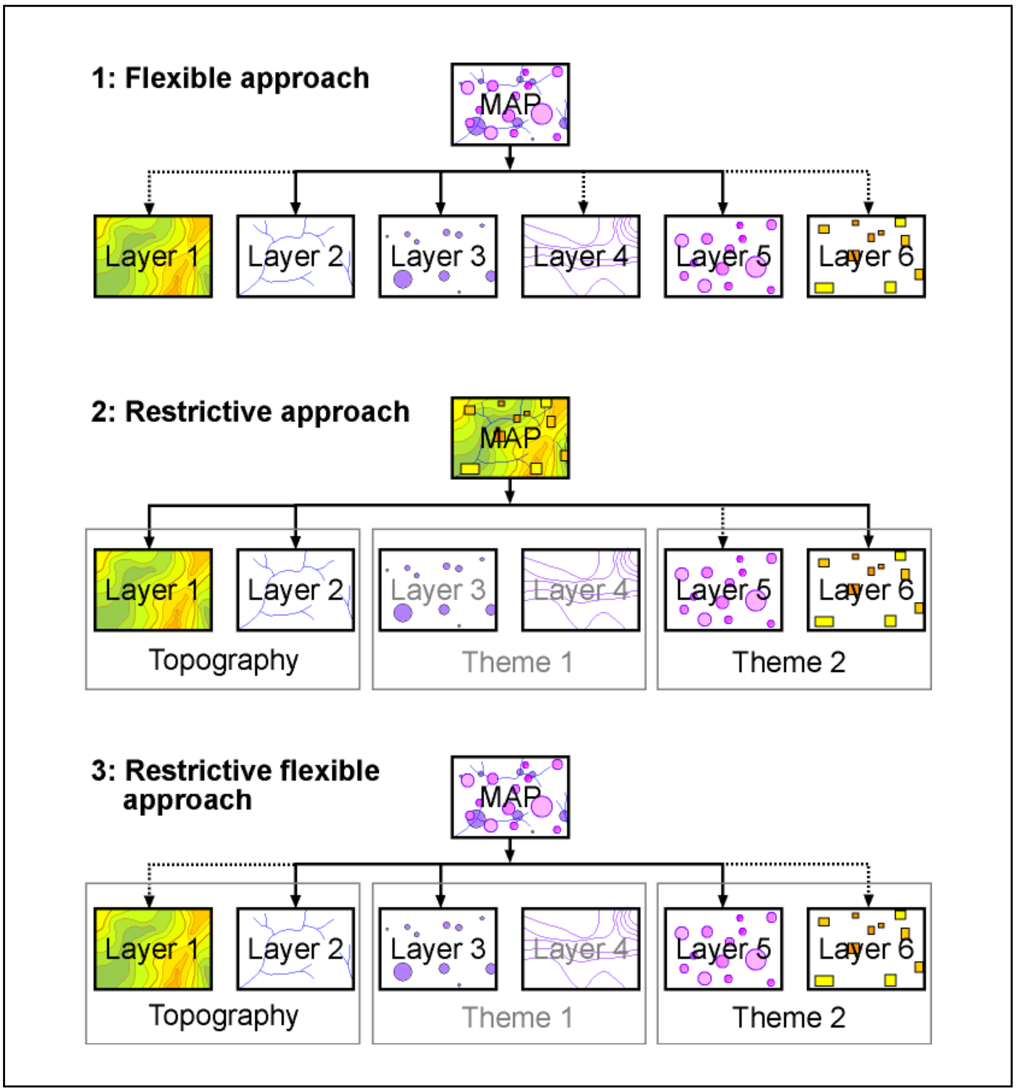

Figure 1: Basic structure of thematic atlas concepts

Data are stored in layers which are composed of geometry, attributes and metadata of certain topics. Layers can be added on demand without destructive impact on the structure. A map is usually built up by a number of such layers (e.g., some topographic layers and some thematic layers).

The **flexible approach** corresponds to a completely open concept. The layers form an unstructured pool and the user is offered to create a map by selecting some of the layers. The user decides which layers to use. If he/she is neither a cartographer nor an expert in the topics to be presented, it may occur easily that maps are produced which make no sense or foster misleading conclusions.

In contrast, the **restrictive approach** corresponds to a static concept. There exists a structure to sort the layers by different themes (topographic base, economy, ecology, population, etc., dependent on scales, map sections, reference dates, data sources, etc.). For the user it is only possible to apply prefabricated maps which consist of a certain number of layers. It is only possible to activate or deactivate the available layers.

A combination of the two strategies is the **restrictive flexible approach**. The data are structured according to topics and themes as with the restrictive approach. But the user has the opportunity to draw additional layers of other themes for further information and auxiliary comparison. In order to ensure reasonable cartographic results it is not possible to combine all existing layers as it is with the flexible approach. It is, e.g., not possible to overlay maps of different scales and aerial sections.

The AtOS concept follows the restrictive flexible approach. In the case of AtOS there exists a hierarchical thematic structure divided into six main theme fields (ecology, population, economy, traffic, spatial and regional planning, miscellaneous). Every theme field consists of several maps. The maps in turn are composed of different layers. The user is able to select a map from each main thematic field. Once a map is selected the user is able to choose several topographic layers, all the thematic layers belonging to the map as well as additional thematic layers from other maps and other main theme fields for comparison.

Another item to be considered is updating. Borders may undergo changes, countries may become independent, others will integrate themselves into confederations, new roads are built etc. As this atlas always will try to present up-to-date information it is not possible to use a uniform data base. However, it is not an easy task to ensure expandable and adjustable data management to meet theses requirements. To this end it is necessary to include time stamps into the data structure and it must be ensured that the user is not in the position to mix up the different data sources. Of course, it might sometimes be interesting to offer a time journey as well.

With the AtOS project flexible data management is assured by a modular structure of the programs and a detailed documentation of data structure and update possibilities. It should be possible to integrate new topics as well as updates of existing ones without changing much of the main structure of the atlas. The main structure of the atlas is illustrated by Figure 2.

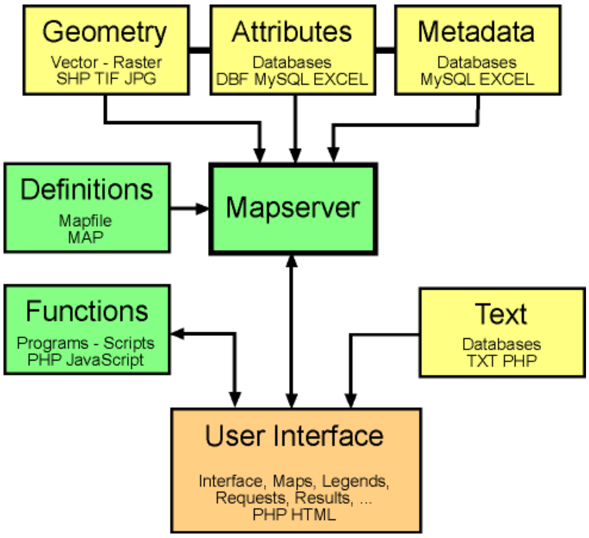

Figure 2: Main structure of the AtOS project

Updates will only affect the areas marked yellow (see Figure 2). The data set of AtOS is separated from the program code (shown in green). If new contents are added as a new map or just as a new layer the following steps have to be performed:

- Collection, harmonization and editing of the data considering the later cartographic presentation (see chapter 1).
- Structuring of the data into geometry, attribute and metadata files.
- Allocation of the layer to the respective map or generation of a new map. Therefore it is necessary to update the Mapserver definitions, functions and the user interface. At this point it is possible to integrate the contents into other maps as layers for comparison.
- Defining the classification, cartographic visualization, the legend, the functions of the layer, etc., in the map file(s).
- Adding the necessary textual information to the text-database (in all languages required).

At the moment AtOS is only in the stage of a prototype and so far only a few map themes are implemented. Also the chapter on this subject must therefore remain incomplete. It was intended just to underline how important an open and flexible concept is for a complex product like this.

3. Preparation of the Internet atlas and its contents for an international auditory

The auditory of a multinational atlas covering a variety of countries and published in the WWW is assumedly an international one. It is necessary to consider this fact with the design of the user interface and with handling of the geographical names.

The user interface has to be multilingual. Therefore the whole textual information (like titles, descriptions, lists, help text etc.) is separated from the files of the program code. All textual information is addressed by variables. The values of the variables are stored in text databases and they can be exchanged easily (see also Figure 2). The user selects the language (for example by clicking on a button) and the interface will be reloaded in the language requested. With this method it is easy to update textual information and it is possible to design the interface in many languages [see also IBM (2000)].

More difficult is the handling of geographical names. There exist a remarkable number of different scripts and languages in the region portrayed by AtOS. The average user from outside or even from the region itself may be acquainted just with a few of them. It is not adequate to use language-specific exonyms in an atlas conceived for an international public. For the sake of scientific correctness, in order to enable secure re-transliteration of transliterated names and not the least for pure courtesy opposite donor countries and languages a correct spelling of all names according to the specifications of the donor alphabets is required. This means the use of all special letters and diacritics of alphabets written in Roman letters as well as the correct rendering of all transliteration systems in the case of donor alphabets written by other than Roman letters [see also Back (1997].

This means, e.g., in the case of Belgrade, the capital of Yugoslavia, that not the Cyrillic endonym (Београд), neither the English (Belgrade) nor the German (Belgrad) exonyms are used on the maps, but the Serbian endonym transliterated into Roman script (Beograd) according to the system of the “Imenik mesta”, the Yugoslavian names gazetteer, which has been recommended for transliteration from Serbian Cyrillic into Roman letters by the United Nations in consent with the donor country.

This means in the case of Bucharest, the capital of Romania, where the Romanian language is written in Roman letters, that the Romanian endonym Bucureşti (and neither the English exonym Bucharest nor the German exonym Bukarest) appears on the maps, although the letter ş is neither used in the English nor in the German alphabets and an English or German reader may doubt how to pronounce it.

However, user access to the data bank and finding of places on the map is not easy for users not acquainted with the endonyms. Therefore also exonyms are offered on the user interface (at the moment exonyms in English and German).

If the endonyms (or transliterations of endonyms) are to be displayed correctly it is necessary to integrate special characters and diacritics, which pose problems in the Internet environment.

3.1. Character and font sets on the Internet

Internally a computer works with byte values. The allocation of byte values to characters is based on character sets. Recently, character sets were defined 8 Bit wide or even smaller. Therefore they can carry 256 different characters at a maximum. Very common were the use of the ASCII or the ANSI character sets (see also Figure 3).

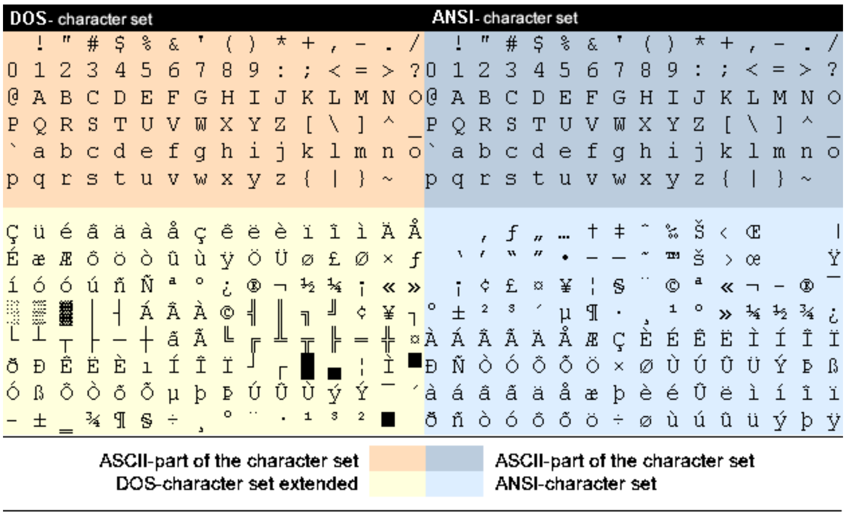

Figure 3: Examples for 8 Bit character sets (DOS and ANSI character set)

The demand for internationally useable character sets has increased in the last few years. A set of character sets became standardized, like the well-known iso-8859-family. There are 16 different character sets defined and they are all supported with HTML 4.0. These character sets meet the requirements of individual script cultures and their alphabets, but they are not flexible enough to be used for international web pages, since not more than 256 different characters can be defined.

For multilingual documents or for work with non-alphabetic scripts it is necessary to use larger character sets. Recently the Unicode Consortium develops the Unicode system. It is a 2 byte character set and can therefore allocate 65536 different characters. Most characters and symbols of the script and symbol systems documented are included into this system. In addition to pure allocation of characters also characteristics of symbols are defined. They define a.o. the direction of typing or transliteration. Many of the modern computer systems (like operating systems and programs) support and make use of the Unicode system. The internationally valid Standard Universal Character Set (UCS) belonging to ISO 10646 is compatible to the Unicode system.

The font set is necessary to visualize the particular characters on the screen. Today only a few font sets exist, which support all the symbols of a 16 Bit character set (e.g. Arial Unicode). It cannot be assumed that the average user has got such a large font set installed on his computer. Therefore it is necessary to provide the fonts on the server side. Although most of the modern browsers support the Unicode character set standard, there may occur some considerable differences between symbols displayed as a result of the different products [see also Münz (1998)].

3.2. Multilingual elements

Since some geographic objects like rivers or mountain ranges extend over more than one language area, data management must be adjusted. To every geographic object a unique object code is attributed. The river Danube, e.g., has been attributed object code number 1423. This river crosses several language areas and assumes therefore several name variants. A name code is added which allows assigning up to 100 name variants to one object. Consequently, the Danube section in Germany and Austria with ID 142300 gets the German name “Donau”, the Slovakian river section with ID 142301 gets the Slovakian name “Dunaj”, etc. This is important for searching the objects. If somebody searches for the geographic object “Dunaj”, he/she should find the whole river, not only the part, where it has this name variant.

3.3. Pronouncing geographical names

The pronunciation of geographical names can be included in form of sound files. They enable the user to click on a name and to listen how it is pronounced in the donor language [more information see Clicksee (2000)].

4. Conclusion

The average Internet user is interested in up-to-date information. Within the AtOS project up-to-date maps are to be published in a temporal and open-ended sequence in the framework of a multinational Internet atlas. This means that the data management concept of the atlas is to be open enough for continuous updates and upgrades. Data from very different sources have to be harmonized and evaluated to achieve a scientifically valid and comparable cartographic presentation – usually in cooperation with experts from the countries represented. Once the data base for a new map is established it should be easy to integrate it into the atlas concept. Although a static concept for data management is not considered appropriate, the Internet requires precise structures and some user guidance to avoid inadequate combinations of layers and misleading conclusions on the user side. With conceiving this Internet atlas it is also important to take into account that map contents as well as auditory are multinational. This has implications on interface design as well as on the rendering of and handling with geographical names.

5. References

- Back, O. (1997): Fragen der Wiedergabe fremdsprachlicher geographischer Namen durch Exonyme oder durch Umschriftung. In: Kretschmer et al (Eds.): Kartographie und Namenstandardisierung. Symposium über geographische Namen. Wien, Institut für Geographie der Universität Wien (= Wiener Schriften zur Geographie und Kartographie; Band 10). Pp. 55-63.
- Clicksee Network Co. (2000): Adding sound into a webpage.
  URL: http://203.146.102.147/sound.asp
- IBM (2000): Introduction to ICU (International Components for Unicode)
  URL: http://oss.software.ibm.com/icu/userguide/index.html
- Jordan, P. (1993): Editing an international map series – specifics and problems. In: Proceedings, Volume 1 of the 16th International Cartographic Conference Cologne/Köln 3-9 May 1993, ed. by Peter Mesenburg, Bielefeld, pp. 273-282.
- Jordan, P. (1994): Die redaktionelle Arbeit an einer internationalen Kartenserie. Besonderheiten und Probleme. In: Kartographische Nachrichten, 44/3, Bonn, pp. 96-104.
- Münz, S. (1998): Selfhtml, Version 7.0
  URL: http://www.teamone.de/selfhtml/selfhtml.htm
- Neumann, A. (1999) – D. Richard: Internet Atlas of Switzerland – New developments and improvements. In: Proceedings of the 19th International Cartographic Conference and the 11th General Assembly of the ICA in Ottawa. August 14 to 21, 1999. Ottawa, CD-ROM, Session 46-D.
- Peterson, M. P. (1999): Trends in Internet Map Use – A Second Look. In: Proceedings of the 19th International Cartographic Conference and the 11th General Assembly of the ICA in Ottawa. August 14 to 21, 1999. Ottawa, CD-ROM, Session 03-C.
- Unicode Consortium (2001): URL: http://www.unicode.org/