POLISH ASSOCIATION FOR SPATIAL INFORMATION
ANNALS OF GEOMATICS 2004 ❍ VOLUME II ❍ NUMBER 1

MULTILINGUAL EUROPEAN SUBSET OF UNICODE IN GEOSPATIAL DATA ENCODING

WIELOJĘZYCZNY EUROPEJSKI PODZBIÓR UNICODE W KODOWANIU DANYCH GEOPRZESTRZENNYCH

Janusz Michalak

Warsaw University, Department of Geology, Warsaw, Poland

Keywords: geospatial data, Unicode, MES, geoinformation system, ESDI, web technology, ISO standards
Słowa kluczowe: dane geoprzestrzenne, Unicode, MES, system geoinformacyjny, ESDI, technologia WWW, standardy ISO

Introduction

Most problems connected with designing and construction of spatial data infrastructure are the same irrespective of the region of the world where this infrastructure is to function. This refers both to its aspects related to technology, architecture and data models and to those related to the users' needs, thematic content elements and methods of access. For this reason, the need to develop worldwide standards in this respect (ISO, 2004) is obvious and does not raise any reservations. However, a part of these problems is of regional character and may be expressed in the international standards only in the form of guidelines and recommendations. Cultural issues of individual countries, including linguistic and legal ones, undoubtedly belong to this group of problems. Europe as a region of the world is featured with great cultural diversity. On the European area of 46 countries there are 205 languages in use (SIL, 2003). Many of these languages have their own alphabets or at least their own national characters extending Latin or Cyrillic alphabet. *The Alphabets of Europe* (Everson, 2002) contains the list of all alphabets used in Europe and there are 165 of them. Contrary to other regions of the world, technological assumptions of European Spatial Data Infrastructure (ESDI) must take this diversity into account. Otherwise, the lack of possibilities of simultaneous and, at the same time, unequivocal encoding and, consequently, of presenting characters of different alphabets will make interpretation of geoinformation contents of the infrastructure resources more difficult or impossible, and finally will significantly restrict its usefulness. In particular, this issue refers to geographical names of small but numerous localities, which in most cases have no equivalents in other languages. As an example, we may make a journey through Europe to a locality

situated close to Jerusalem (ירושלים), with the following main points: we start in Gråträsk (town in Sweden) and pass the following localities: Светлогорск (Russian town in the Kaliningrad District), Zbąszyń, Żagań and Łódź (cities in Poland), a locality close to Athens (Αθήα) and Keçiören (town in Turkey). Without the use of Unicode or other standards extending the character set recording of these names (in ASCII code) would look as follows: ”□□□□□□□”, ”Gr□tr□sk”, ”□□□□□□□□□□□”, ”Zb□szy□”, ”□aga□”, ”□□d□”, ”□□□□□”, ”Ke□i□ren”.

The authors of the concept of European Spatial Data Infrastructure (ESDI), which was planned and developed within the INSPIRE (Infrastructure for Spatial Information in Europe) initiative assign proper weight to the multilingual problems of European societies. A good example provides a fragment of the document entitled *INSPIRE Architecture and Standards Position Paper* (INSPIRE-AST WG, 2002): *Multilingual aspects relate to almost all functionality envisaged. They concern the querying of metadata, viewing of GI (place names, labels), and the results of any analysis (e.g., querying a data set). Multi-lingual support to INSPIRE is therefore imperative.*

For these reasons, the author conducted analysis of possibilities to use Unicode in ESDI with particular attention paid to geographical names. Among others, a test experiment was performed related to finding information connected with geographical names recorded in Unicode on a map placed on website by means of Google searcher.

What is Unicode?

The website of Unicode Consortium (UC, 2004) is a source of exhaustive information on Unicode, among others with respect to programming environments and implementation platforms meeting this standard, and an example of using Unicode in web technology. The website also contains the list of application software systems conformant with this standard, but this list does not include software for GIS and geoinformation infrastructures. On this website, the answer to the question put in the subtitle may be found: *Fundamentally, computers just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding could contain enough characters: for example, the European Union alone requires several different encodings to cover all its languages. Even for a single language like English no single encoding was adequate for all the letters, punctuation, and technical symbols in common use. These encoding systems also conflict with one another. That is, two encodings can use the same number for two* different *characters, or use different numbers for the* same *character. Any given computer (especially servers) needs to support many different encodings; yet whenever data is passed between different encodings or platforms, that data always runs the risk of corruption. Unicode is changing all that! Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language*. (UC, 2004).

ISO standards of text encoding in geoinformation

The standard ISO 19118: Geographic information – Encoding (ISO, 2002) defines, among others, the rules of encoding texts contained in geospatial data sets. Most often these texts appear as indirect references of features, that is by means of geographical names, which may be changed into coordinates; also as descriptions of features or values of textual attributes characterising these features or associations between them. Geographical names have the most meaningful role in these cases, whose importance was indicated in the example presented in the introduction.

Character encoding standards generally used up till quite recently such as ACII (7-byte) and ISO 5988-1 to 5988-15 (8-byte) proved to be insufficient for textual encoding in different languages without any conflicts (Bień, 1998). For this reason, according to the standard ISO 19118, texts contained in geoinformation should be encoded in one of two variants of code tables defined in the standard ISO/IEC 10646 (ISO, 1999):

1. UCS-4 (Universal Character Set in 4 octets) – 31-byte set of characters for all possible languages in the world, also taking into account languages, which are important for scientific reasons only, for instance hieroglyphs. In the terminology used in this standard, an octet is a two digit hexadecimal number occupying place of one byte in the memory (corresponding to eight-digit binary number). As in this case any character is encoded by means of 4 octets, and in practice only two of them are used, the variant described below is much more often used.
2. UCS-2 (ISO/IEC 10646-1, corresponding to industry standard Unicode) – Subset UCS-4 restricted to possibilities provided by encoding by means of 2 octets and called ”Basic Multilingual Plane” (BMP) or ”Plane 0”. It covers all languages based on Latin alphabet, and also on Greek, Hebrew, Cyrillic, Arabic, Korean Hongul, Japanese Katakana alphabets and many others.

Characters defined in code tables (UCS-4 and UCS-2) may be encoded according to a few rules divided into two groups:

1. With fixed size of code of single character:
   1.1. For UCS-4 sequences of 4-byte codes.
   1.2. For UCS-2 sequences of 2-byte codes.
2. With variable size of code of single character (UTF – UCS Transfer Format):
   2.1. UTF-8 – characters from the range 0x00 to 0x7F (corresponding to ASCII code) are encoded as 1-byte numbers. Other characters (from the range above 0x7F) are encoded by means of sequences of 2 to 3 bytes for codes from the table UCS-2 and up to 6 bytes for codes from the table UCS-4. Table 1 presents the UTF-8 coding format.
   2.2. UTF-16 – characters from the range 0x0000 to 0xFFFF (contained in the table UCS-2) are encoded as 2-byte characters. Characters from the range from 0x10000 to 0x10FFFF (contained in the table UCS-4) are encoded by means of two 4-figure hexadecimal numbers (16-figure binary numbers): the first covers the range from 0xD800 to 0xDBff, and the second from 0xDC00 to 0xDFFF. Characters with codes above 0x10FFFF cannot be coded in UTF-16 format.

Table 1. UTF-8 byte sequences to represent a character (ISO, 2002)

Code Range: (in hexadecimal numbers)	**Format of byte sequence:** (0 to 111111 – fixed bits, xxxx – bit positions filled with the bits of the character code number in binary representation)
0x00000000 – x0000007F (ASCII range)	0xxxxxxx
0x00000080 – 0x000007FF (UCS-2 range)	110xxxxx 10xxxxxx
0x00000800 – 0x0000FFFF (UCS-2 range)	1110xxxx 10xxxxxx 10xxxxxx
0x00010000 – 0x001FFFFF (UCS-4 range)	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
0x00200000 – 0x03FFFFFF (UCS-4 range)	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
0x04000000 – 0x7FFFFFFF (UCS-4 range)	1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

MES – Multilingual European Subset of Unicode

Multilingual European Subset (MES) was defined by a team established by CEN (European Committee for Standardization) in 1998 and this subset is described in the document CWA 13873:2000 (CEN, 2000). The task of the team was to determine a standard set of characters (letters, punctuation marks and symbols), which are used in computer systems for the needs of administration and business in Europe. It was decided that using of the full set of characters defined in the standard ISO 10646-1 is not rational for European needs both for technical and economic reasons. In the results of works and negotiations, three MES subsets were created for different applications:

- MES-1: the set for alphabets derived from the Latin alphabet and based on the standard ISO/IEC 6937:1994. This is a fixed set containing 334 characters from *Latin-1* to *Latin-5* repositories, and the mark of the currency euro. This set covers 44 languages of the European area.
- MES-2: the set of alphabets derived from Latin, Greek and Cyrillic alphabets based on the standard ENV 1973:1996. This is also a fixed set containing 1062 characters for more than 128 languages.
- MES-3: the set comprising all characters of all alphabets of the area of Europe. This set is defined in two versions: the fixed one (MES-3B) and non-fixed one (MES-3A). MES-3B as a fixed set comprises 2819 determined characters, and the initial composition of the set MES-3A is identical, but may be extended in the future. The set MES-3, and also its subsets: MES-1 i MES-2 comprise languages, in which you write from left to right and thus do not include Hebrew. But the open formula of the version MES-3A allows introducing this extension.

Unicode in web technologies concerning geoinformation

Most web browsers are able to handle Unicode characters in the UTF-8 format. For this reason, this method of encoding seems to be the most appropriate for applications, which make geoinformation available by internet. A good example of such a solution provides GNS (GEOnet Name Server), which task is, among others, to changed geographical names given in w Unicode (UTF-8) into geographic coordinates (NGA, 2003; 2004). This server contains nearly 5.5 million names of geographic features from the whole world. The example of the result of search for the name ”Warszawa” is given in Fig. 1.

Web searchers such as, for instance Google, also can handle Unicode characters encoded in the UTF-8 format. The author conducted a test experiment on connecting web searchers capabilities with the application of extensions of HTML language to encoding of geographical names (as a separate textual layer in the meaning of HTML language) on a map sent from a web browser. The test page with the Polish version may be found at http://netgis.geo.uw.edu.pl/unicode/index.shtml and the English version at http://netgis.geo.uw.edu.pl/unicode/index-en.shtml. This experiment proved full usefulness of combining webmapping technology with searching systems for connecting geographical names encoded on maps in Unicode with

REGION	DESIG.	NAME / LATITUDE	LONGITUDE	AREA	UTM	JOG NO.	UFI
		Make sure your browser is set to UNICODE (UTF-8) encoding					
Native		Warszawa Wschodnia					
(2)	RSTN	52° 15' 00" N	021° 03' 00" E	PL67	EC08	NN34-11	-534437
Native		Warszawa Główna					
(2)	RSTN	52° 14' 00" N	021° 00' 00" E	PL67	EC08	NN34-11	-534436
Variant		Warsaw Main Station (UNI= -757510)					
Native		Warszawa Gdańska					
(2)	RSTN	52° 16' 00" N	021° 00' 00" E	PL67	EC09	NN34-11	-534435
Variant		Warsaw-Gdanska Station (UNI= -757509)					
Variant		Gdański, Dworzec (UNI= -714326)					
Native		Warszawa Dworzec Zachodni					
(2)	RSTN	52° 14' 00" N	020° 59' 00" E	PL67	DC98	NN34-11	-534434
Variant		Zachodni, Dworzec (UNI= -761858)					
Variant		Warsaw-West Passenger Station (UNI= -757511)					

Fig. 1. A fragment of webpage sent by GEOnet Name Server (GNS) in response to a query related to the geographical name ”Warszawa” (NGA, 2004).

Fig. 2. A fragment of webpage containing window of Google searcher and a map with additional textual layer (encoded independently in the meaning of HTML language), on which geographical names are placed in Unicode. Copying of the name from the map to the searcher window and using the search command according to the instruction contained in Table 2 gives the result presented in Fig. 3.

webpages containing these names. The course of the test is presented in Fig 2 and Fig. 3 and description of actions is contained in Table 2. Simultaneous search of two names which are on the map: Άθήνα and ירושלים (Athens and Jerusalem) gives as a result only two pages with the addresses given above, because searching system of Google has found only these two pages containing both names at the same time.

Fig 3. Webpage search result of two geographical names: Άθήνα and ירושלים (Athens and Jerusalem) copied from the map in Fig. 2. The only two webpages containing both names at the same time are those connected with this experiment. A fragment of the other page is presented in Fig. 2.

Table 2. Instruction for conducting the test

Sequence of actions for the page presented in Fig 2	Result
1. Move cursor to the selected geographical name and double click with left key of the mouse (for names composed of two words you should click three times)	The name is highlighted (white letters against black background)
2. Press keys CTRL+C	The name is copied
3. Move cursor to the input text box of the searcher and click with the left key of the mouse	In the left side of the box a vertical blinking dash appears
4. Press keys CTRL+V	The copied name is entered in the box
5. Move cursor to the key "Search in Google" and click with left key of the mouse	New page of Google appears showing the search result (Fig. 3)

For a fictitious name of a locality „Koziabroda Wielka" the test also gives in the result only these two pages. The latter case allows verification of geographical names with respect to their authenticity or correctness of encoding.

Obstacles in applying Unicode in geoinformation

The possibility to use Unicode and its MES subset in geoinformation systems depends on technology, implementation platforms and operational systems on which the geoinformation systems under consideration are based. Geoinformation systems with older solutions encounter serious implementation problems, but Unicode service is an integral part of more recent solutions. In the latter case, construction of multilingual applications is much simpler.

C language is an example of the first group with still used single-byte basic types of textual data: the char for single character and C-string (zero-terminated character-string) specified by the pointer char * to the table (to the string) for sequence of characters. For this reason, various practical solutions are used for encoding texts in Unicode which avoid this problem. For instance, in the library ClibPDF for this language (for processing documents in PDF format) encoding of strings of two-digit hexadecimal numbers is used corresponding to individual 16-byte UCS-2 codes in 8-byte tables (strings) pointed by variables of the char * type:

```
char *annot_title = „FEFF65E5672C8A9E306E6CE891C8306E4F8B”;
```

Another solution used in software developed in the C language provides variables of wchar_t type and pointers to the tables of this type. However, because of implementation differences between various different C compilers this method cannot be treated as standard. Standardisation efforts to extend language C to be able to handle Unicode are still under way (Task Force ISO/IEC JTC1/SC22/WG14) (ISO, 2003a) and their preliminary results are contained in Work Technical Report (ISO, 2003b), which introduces two new simple types: char16_t and char32_t for the needs of Unicode.

At present, when we want to correctly solve the problem of Unicode in languages C and C++, also in cooperation with applications in Java, we may take advantage of the ICU libraries (International Components for Unicode) with Open Source status. ICU is a widely used set of C/C++ and Java libraries to support Unicode and software internationalization and globalization. These libraries are developed within the framework of projects sponsored by IBM. An example of using ICU components for coding names of domains according to IDNA (Internationalizing Domain Names in Applications) standard is presented in Fig. 4.

Comments on the use of Unicode by GIS software producers are given below:

- Core software modules of Arc/Info and ArcMap (ESRI) do not handle Unicode, because their basic programming language is C. For this reason, in one application only one set of characters with the range of one octet (up to 255 characters) may be used. For Far Eastern languages special versions of this software are designed, which cannot be applied to European languages.
- Geomedia software of Intergraph functions only in the environment of Microsoft Windows operational system and, therefore, uses Character Maper (charmap.exe) program for characters other than those in the set of *Latin 1*.
- In GIS GRASS (Open Source) the problem of characters other than the basic set of English alphabet (ASCII – 7 bit) is not solved.
- Oracle DBMS, often used for geospatial databases, has the capability to handle Unicode in UTF-8 format.
- Majority of new software to make webmapping available is developed in Java and, for this reason, have the capability, at least theoretical, to handle Unicode.

The problem of handling texts in Unicode in systems designed for interoperability in geoinformation infrastructures is only casually mentioned or completely passed over in the available literature. For this reason, strenuous tests are required to fully investigate how individual elements of the infrastructure can handle several European languages at the same time.

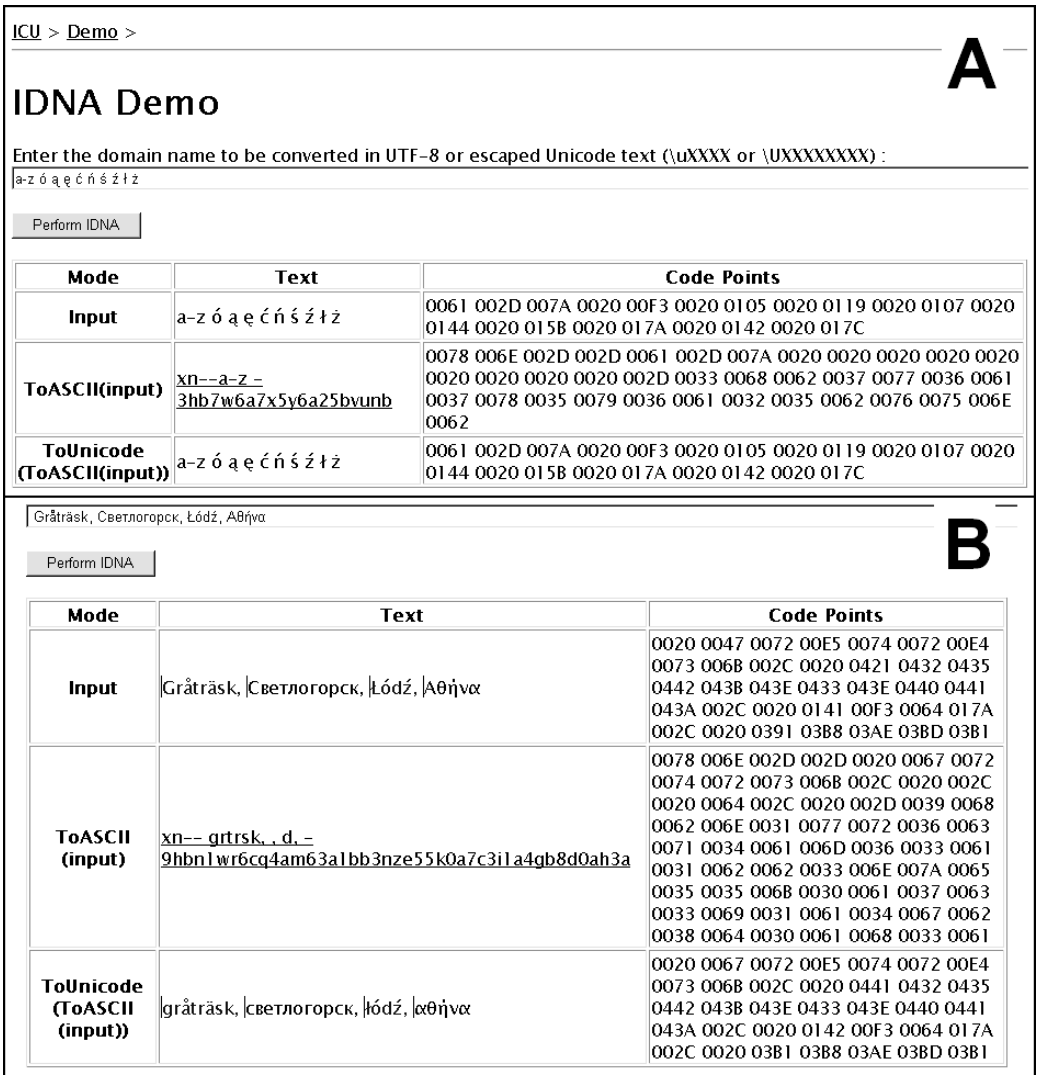

Fig 4. A fragment of the searcher window presenting conversion of Polish letters (example A) from Unicode into ASCII and the opposite from ASCII into Unicode. This convertor is designed to exchange names in internet domains, but it may convert geographical names as well (example B): Gråträsk, Светлогорск, Łódź and Αθήνα.

Conclusions

The results of studies and experiments performed and presented here may be expressed in the form of the following conclusions:

- European Geospatial Data Infrastructure requires application of technological solutions, which allow using all alphabets used on the area of Europe.
- Rational solution of this problem should be based on Unicode standard with character repertoire limited to MES subset.
- Technological nature of geoinformation infrastructure requires separation of external representation of data (in the case of sending them from one system to another) from internal representation (storage and processing within these systems). In the first case, the requirements are determined by the standard ISO 19118. In the second case, encoding of texts does not have to follow the same strict requirements. However, for their effective interoperability it is advisable to use within the system the same method of encoding as outside the system.
- For external representation, more and more often now presented by means of XML language application, the most appropriate method of encoding is UTF-8.
- In many cases software systems based on older development environments are used as elements of geoinformnation infrastructure, for instance written in C language without the possibility of direct handling of Unicode. In such a situation, systems of local or national level processing geoinformation with texts within one language group (in the meaning adopted in the standard ISO 8859), may internally use one-byte (with the range of one octet) methods of character encoding defined in the standard ISI 8859.
- In the situation, where internal encoding method is different than those defined in the standard ISO 19118, interfaces linking these systems with external elements of infrastructure must have capability to convert codes in both directions.
- Software systems for network applications, generally used at present, for instance web servers, web browsers and searchers, can handle various variants of Unicode and various formats of encoding, and at least UTF-8.
- Linking these capabilities with technologies based on standards ISO/TC 211 and on specifications Open GIS Consortium concerning geoinformation will allow to effectively solve the problem of multilingual geographical names in geoinformation infrastructure presented in this paper.

References

Bień J. S., 1998: *Kodowanie tekstów polskich w systemach komputerowych*. Postscriptum, nr 27/29. URL: http://sjikp.us.edu.pl/ps/ps_29_04.html.

CEN, 2000: *Multilingual European Subset in ISO/IEC 10646-1*. CEN Workshop Agreement – CWA 13873:2000. URL: http://www.evertype.com/standards/iso10646/pdf/cwa13873.pdf.

Everson M., 2002: *The Alphabets of Europe*. Version 3.0. URL: http://www.evertype.com/ alphabets/ index.html

IBM, 2004: *International Components for Unicode*. URL: http://oss.software.ibm.com/icu/.

INSPIRE-AST WG, 2002: *INSPIRE Architecture and Standards Position Papers*. JRC – Institute for Environment and Sustainability, Ispra. URL: http://inspire.jrc.it/reports/position_papers/ inspire_ast_pp_v4_3_en.pdf.

ISO, 2003a: *Business Plan and Convener's Report*. Document N1014 of Working Group ISO/IEC JTC1/ SC22/WG14. URL: http://wwwold.dkuug.dk/jtc1/sc22/open/n3607.pdf.

ISO, 2004: *Draft Business Plan of ISO/TC 211 - Geographic information/Geomatics*. ISO/TC 211 Document 1297. URL: http://www.isotc211.org/opendoc/211n1297/211n1297.doc

ISO, 2002: *Geographic information – Encoding*. Draft International Standard ISO/DIS 19118. URL: http://www.isotc211.org/protdoc/DIS/ISO_DIS_19118_(E).pdf.

ISO, 2003b: *Information Technology – Programming languages, their environments and system software interfaces – Extensions for the programming language C to support new character data types*. Document N1040 of Working Group ISO/IEC JTC1 SC22 WG14. URL: std.dkuug.dk/ JTC1/SC22/WG14/www/docs/n1040.pdf

ISO, 1999: *ISO 10646-1: Universal Multiple-Octet Coded Character Set (UCS)*. Second Edition text, Draft 2. ISO/IEC JTC1/SC2/WG2. URL: http://std.dkuug.dk/JTC1/SC2/WG2/docs/ n2005/n2005.pdf.

NGA, 2004: *GEOnet Name Server (GNS)*. National Geospatial-Intelligence Agency. URL: http://earth-info.nima.mil/gns/html/index.html.

NGA, 2003: *GNS-to-UNICODE Character Mapping Tables*. GEOnet Name Server, National Geospatial-Intelligence Agency. URL: http://earth-info.nima.mil/gns/html/gns_faq/gns2unicode /gns2unicode.html.

SIL, 2003: *Languages of Europe.* In: Ethnologue country index – Languages of the world. 14th Edition. URL: http://www.ethnologue.com/country_index.asp?place=Europe.

UC, 2004: *What is Unicode?* Unicode Consortium. URL: http://www.unicode.org/standard/WhatIsUnicode.html.

STRESZCZENIE

*Większość problemów dotyczących infrastruktur geoinformacyjnych w zakresie technologii, architektury i modeli danych jest taka sama, niezależnie od regionu świata, dla którego infrastruktura jest budowana. Z tego powodu potrzeba opracowania ogólnoświatowych standardów w tym zakresie (ISO, 2004) jest faktem, który nie budzi niczyich zastrzeżeń. Jednak część tych problemów ma charakter regionalny i w międzynarodowych standardach może być ujęta jedynie w formie zaleceń. Do tej grupy należą zagadnienia kulturowe poszczególnych krajów, a w tym językowe i prawne. Europa wyróżnia się szczególną różnorodnością kulturową. W 46 krajach europejskich jest obecnie używanych ponad 205 języków (SIL, 2003). Publikacja Eversona (Everson, 2002) zawiera zbiór 165 alfabetów używanych w Europie. Założenia technologiczne Europejskiej Infrastruktury Geoinformacyjnej (ESDI) muszą uwzględniać tą różnorodność, ponieważ brak możliwości jednoczesnego i zarazem jednoznacznego kodowania znaków różnych europejskich alfabetów utrudni interpretacje geoinformacji i przez to znacznie ograniczy jej użyteczność. W szczególności zagadnienie to dotyczy nazw geograficznych małych, ale za to bardzo licznych miejscowości, które nie mają swoich odpowiedników w innych językach. Przykładem może być podróż przez Europę ze szwedzkiego miasta Gråtråsk do miejscowości położonej w pobliżu Jerozolimy (*ירושלים*, której głównymi punktami są miejscowości: Светлогорск (miasto w Okręgu Kaliningradzkim), Zbąszyń, Żagań i Łódź (miasta w Polsce), miejscowość w pobliżu Aten (Ἀθήνα) i Keçiören (miasto w Turcji).*

Autorzy podstaw koncepcji ESDI rozwijanej w inicjatywie INSPIRE doceniają wagę problemów wielojęzyczności społeczeństw europejskich. Przykładem tego jest fragment jednego z dokumentów inicjatywy INSPIRE (INSPIRE-AST WG, 2002): „Aspekty wielojęzyczności są związane z prawie wszystkimi zagadnieniami funkcjonalności. Dotyczą one zapytań o metadane, przeglądania informacji geoprzestrzennej (nazwy miejscowości, objaśnień) i wyników wszelkich analiz (np. wyszukiwania zbiorów danych). Uwzględnienie wielojęzyczności w projekcie INSPIRE jest koniecznością."

Z tych powodów autor przeprowadził analizę możliwości zastosowania Unicode w ESDI ze szczególnym uwzględnieniem nazw geograficznych. Między innymi został przeprowadzony eksperyment testowy dotyczący znajdowania przy pomocy wyszukiwarki Google informacji związanych z nazwami geograficznych zapisanych w Unicode na mapie umieszczonej na stronie WWW.

Czym jest Unicode?

Każdy znak występujący w tekście (litera, symbol lub znak interpunkcyjny) w systemie informatycznym jest reprezentowany przez przyporządkowaną mu liczbę. W rezultacie nazwa, napis lub tekst jest ciągiem liczb. Podstawowy problem polega na długości miejsca, jakie dla pojedynczej liczby jest przeznaczone w pamięci lub nośniku danych. W przypadku jednego lub kilku języków europejskich wystarcz jeden bajt, co odpowiada liczbom z zakresu 0 do 255. Oparty na tym standard ISO 8859 specyfikuje szereg zestawów kodów dla różnych grup języków, lecz stosowanie go nie pozwala na jednoczesne używanie w jednym tekście znaków pochodzących z alfabetów należących do różnych grup (Bień, 1998). Problem ten rozwiązuje standard ISO 10646 (ISO, 1999), popularnie nazywany Unicode, określający reguły kodowania znaków przy pomocy liczb z zakresu 2 lub 4 oktetów, czyli o stałej lub zmiennej długości od jednego do czterech bajtów. Witryna Unicode Consortium (UC, 2004) jest źródłem wyczerpujących informacji na temat Unicode.

Standardy ISO dotyczące kodowania tekstów w geoinformacji

Norma ISO 19118 (ISO, 2002) specyfikuje między innymi reguły kodowania tekstów zawartych w zestawach danych geoprzestrzennych. Najczęściej teksty te wstępują jako odniesienie pośrednie wyróżnień, opisy wyróżnień lub wartości atrybutów tekstowych charakteryzujących te wyróżnienia. Powszechnie stosowane do niedawna sposoby kodowania znaków, takie jak ACII (7-bitowy) i ISO 5988-1 do 5988-15 (8-bitowe) okazały się niewystarczające do bezkonfliktowego, z punktu widzenia różnych języków, zapisu informacji geoprzestrzennej. Z tego względu, według standardu ISO 19118, teksty zawarte w geoinformacji powinny być zapisywane w jednym z dwóch wariantów tablic kodowych zdefiniowanych w standardzie ISO 10646 (ISO, 1999):

1. *UCS-4 (*Universal Character Set in 4 octets*) – 31-bitowy zbiór znaków dla wszystkich możliwych języków na świecie.*
2. *UCS-2 (ISO/IEC 10646-1) - Podzbiór UCS-4 ograniczony do możliwości, jakie daje kodowanie przy pomocy 2 oktetów i nazywany* Basic Multilingual Plane *(BMP). Obejmuje on wszystkie języki bazujące na alfabecie łacińskim, a także alfabety: grecki, hebrajski, cyrylicę, arabski i alfabety dalekowschodnie.*

Znaki zdefiniowane w tablicach kodowych UCS-4 i UCS-2 mogą być zapisywane według kilku reguł dzielących się na dwie grupy:

1. *O stałej długości kodu jednego znaku: dla UCS-4 ciągi kodów 4-bajtowych i dla UCS-2 ciągi kodów 2-bajtowych.*
2. *O zmiennej długości kodu jednego znaku (UTF –* UCS Transfer Format*):*
   UTF-8 – znaki z zakresu 0x00 do 0x7F są zapisywane jako 1-bajtowe. Pozostałe znaki są zapisywane przy pomocy sekwencji od 2 do 3 bajtów dla kodów UCS-2 i do 6 bajtów dla kodów UCS-4.
   UTF-16 – znaki z zakresu 0x0000 do 0xFFFF są zapisywane jako 2-bajtowe. Znaki z zakresu od 0x10000 do 0x10FFFF są zapisywane przy pomocy dwóch 16-bitowych liczb całkowitych.

MES – „Multilingual European Subset" dla Unicode

MES został zdefiniowany przez zespół powołany przez CEN i jest opisany w dokumencie CWA 13873:2000 (CEN, 2000). Uznano, że stosowanie dla potrzeb europejskich pełnego zestawu znaków zdefiniowanych w normie ISO 10646-1 nie jest racjonalne ze względów technicznych i ekonomicznych. W rezultacie wykonanych prac i uzgodnień powstały trzy podzbiory MES dla różnych zastosowań:

- *MES-1: zestaw dla alfabetów wywodzących się z alfabetu łacińskiego. Jest to zamknięty zbiór zawierający 334 znaki ze zbiorów* Latin-1 *do 5 i obejmuje 44 europejskie języki.*

- *MES-2: zestaw dla alfabetów wywodzących się z alfabetu łacińskiego, greckiego i cyrylicy. Ten zestaw jest również zamkniętym zbiorem i zawiera 1062 znaki dla ponad 128 języków.*
- *MES-3: zestaw obejmujący wszystkie znaki wszystkich alfabetów obszaru Europy. Zestaw ten jest zdefiniowany w dwóch wersjach: zamkniętej (MES-3B) i niezamkniętej (rozszerzalnej) (MES-3A). MES-3B, jako zbór zamknięty, obejmuje 2819 ustalonych znaków, a początkowy zestaw zbioru MES-3A jest identyczny.*

Unicode w technologiach WWW dotyczących geoinformacji

Większość przeglądarek WWW potrafi interpretować znaki Unicode w formacie UTF-8. Z tego względu ten sposób zapisu wydaje się być najbardziej odpowiedni dla zastosowań udostępniania geoinformacji przez WWW. Przykładem takiego rozwiązania jest GEOnet Name Server (NGA, 2003; 2004). Większość wyszukiwarek internetowych, jak na przykład Google, również obsługuje Unicode. Powiązanie ich możliwości z zastosowaniem rozszerzeń języka HTML do zapisu nazw geograficznych (jako oddzielnej warstwy tekstowej) na mapie przesyłanej do przeglądarki WWW było przedmiotem eksperymentu testowego wykonanego przez autora i dostępnego pod adresem http://netgis.geo.uw.edu.pl/unicode/index.shtml. Test ten dla fikcyjnej nazwy nieistniejącej miejscowości „Koziabroda Wielka" daje w wyniku tylko dwie strony znajdujące się w tym serwerze i związane z testem.

Przeszkody w stosowaniu Unicode w geoinformacji

Możliwość stosowania Unicode i jego podzbioru MES w systemach geoinformacyjnych jest uzależniona od technologii, platform implementacyjnych i systemów operacyjnych, na których te systemy są oparte. Można tu dokonać podziału na rozwiązania starsze, w których napotyka się na istotne problemy implementacyjne i nowsze, w których obsługa Unicode jest ich integralną częścią, co znacznie ułatwia budowanie aplikacji wielojęzycznych.
*Przykładem pierwszej grupy jest język C i ciągle jeszcze powszechnie stosowane w nim jednobajtowe typy podstawowych danych tekstowych: znak char i łańcuch znaków (*C-string*) jako wskaźnik do tablicy znaków char *. Z tego powodu dla zapisu tekstów w Unicode w języku C stosowane są różne praktyczne rozwiązania, na przykład zapis łańcuchów liczb heksadecymalnych odpowiadających poszczególnym 16-bitowym kodom UCS-2 w 8-bitowych zmiennych typu char *. Prace standaryzacyjne nad rozszerzeniami języka C dla obsługi Unicode jeszcze trwają (ISO, 2003a)i są zawarte w roboczym raporcie technicznym ISO/IEC DTR 19769 (ISO, 2003b).*

Unicode w najczęściej stosowanych systemach programowych dla geoinformacji

- *podstawowe moduły programowe ESRI (Arc/Info, ArcMap) nie obsługują Unicode, ponieważ ich językiem programowania jest podstawowe C. Dla języków dalekowschodnich przeznaczone są specjalne wersje tego oprogramowania oparte na niestandardowych rozwiązaniach.*
- *Geomedia firmy Intergraph działa tylko w środowisku systemu operacyjnego Microsoft i z tego względu wykorzystuje systemowy program Character Maper.*
- *oprogramowanie Open Source GIS GRASS nie ma rozwiązanego problemu znaków innych niż podstawowy zestaw alfabetu angielskiego.*
- *oprogramowanie SZRBD Oracle, jako baza danych geoprzestrzennych ma możliwość obsługi Unicode w formacie UTF-8.*
- *większość nowego oprogramowania dla udostępniania map w WWW jest opracowane w języku Java i z tego względu ma, przynajmniej teoretyczną, możliwość obsługi Unicode.*

Zagadnienie obsługi Unicode w szeroko obecnie stosowanych systemach geoinformacyjnych wymaga żmudnych prac testowych, ponieważ w dostępnej na ten temat literaturze jest poruszane jedynie zdawkowo i zupełnie pomijane.

Wnioski

Rezultat przedstawionych tu prac studialnych i eksperymentalnych można ująć w formie następujących wniosków:

- *ESDI wymaga zastosowania rozwiązań technologicznych pozwalających na stosowanie wszystkich alfabetów, jakie są używane na obszarze Europy. Racjonalne rozwiązanie tego problemu powinno być oparte na standardzie Unicode z ograniczeniem do podzbioru MES.*
- *Charakter technologiczny infrastruktury geoinformacyjnej wymaga oddzielenia zewnętrznej reprezentacji danych od reprezentacji wewnętrznej, która nie musi podlegać takim samym ostrym wymaganiom. Dla reprezentacji zewnętrznej, obecnie coraz częściej przedstawianej przy pomocy aplikacji języka XML, najbardziej odpowiednim sposobem kodowania jest UTF-8.*
- *W wielu przypadkach, jako składniki infrastruktury geoinformacyjnej, używane są systemy programowe oparte na starszych środowiskach narzędziowych, na przykład języku C nieobsługującym bezpośrednio Unicode. Z tego względu systemy o zasięgu lokalnym lub narodowym mogą wewnętrznie posługiwać się sposobami kodowania znaków określonymi w standardzie ISO 8859. W takich przypadkach interfejsy łączące te systemy z zewnętrznymi elementami infrastruktury muszą mieć możliwość dokonywania konwersji kodu w obu kierunkach.*
- *Powszechnie obecnie stosowane systemy programowe dla zastosowań sieciowych, na przykład serwery, przeglądarki i wyszukiwarki WWW, na ogół potrafią obsługiwać różne warianty Unicode i różne formaty zapisu tego kodu, a przynajmniej format UTF-8. Powiązanie tych możliwości z technologiami opartymi na standardach ISO/TC 211 i specyfikacjach Open GIS Consortium dotyczącymi geoinformacji pozwoli skutecznie rozwiązać przedstawiony tu problem nazw geograficznych zapisanych w różnych językach.*

Janusz Michalak
J.Michalak@geo.uw.edu.pl
http://geo.uw.edu.pl
http://testbed.ptip.org.pl