MULTILINGUAL-MULTI SCRIPT PLACE NAME INFORMATION SYSTEM FOR WEB MAPPING IN XINJIANG CHINA

Alishir Kurban[1], Xi Chen[1], Abdimijit Ablimit[1] and Kristien Ooms[2]
[1]Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences
No. 818, South Beijing Road, Urumqi, Xinjiang 830011, China
alishir@ms.xjb.ac.cn

[2]Ghent University,
Department of Geography
Krijgslaan 281, S8, B-9000, Ghent, Belgium

Con formato: Neerlandés (Bélgica)

ABSTRACT

Con formato: Neerlandés (Bélgica)

Since Xinjian is a very attractive place in China, it is essential that the people visiting it (tourists, businessmen, etc.) can obtain accurate and detailed information about it in its local language(s). The multilingual-Multi Script place name database – targeting both global and local users – is constructed to satisfy this need. This paper describes the development of a Multilingual-Multi Script Place Name Information System for Web Mapping in Xinjiang, with the aim of creating a detailed regional map service through the Internet. Describing the regional geographical information in its local language and subsequently sharing this data world wide demands for elaborate testing of each element of the web service for all the languages and writing scripts used in that region; Uyghur, Chinese, Russian and English in this case.

In addition to the different writing scripts, the website also allows to obtain the correct pronunciation of a name in one of the local language to improve the communication in the region. Voice recordings of the local names in Uyghur and Chinese are stored in a multimedia database and linked to the website to achieve this service. Currently, the website is still in the prototype-phase and is tested extensively to ensure and improve its flexibility and stability.

1. Introduction

Since the current economy becomes more and more global, interactions between different cultures are bound to happen more regularly. Different languages and writing scripts are inherently linked with these cultures, influencing the communication worldwide. Multilingualism is a widespread phenomenon as there are more than 6,800 languages across the globe. Furthermore, there is no

one-to-one relation between a language and a region: multiple languages can be used in a single region, but a number of languages - such as English, Spanish and French - are also (officially) used in several different countries in the world. (Deckert, 2004; Veselionova and Booza, 2006)

Con formato: Sin Resaltar

The first situation is especially true for the Xinjiang Uyghur Autonomous Region (XUAR), which is the largest province in China, covering 1/6th of the country's surface. The region has borders with a rather large number of countries: Mongolia, Russia, Kazakhstan, Kyrgyzstan, Tajikistan, Afghanistan, Pakistan and India. There are 13 minorities, which include Uyghur, Kazak, Kyrgyz, and Uzbek, resulting in many ethnic languages that use their own writing scripts. These scripts originate from Arabic characters but where modified over time. Since the XUAR is a very attractive place for traveling, trading and investing, it is essential that the customers are provided with complete and detailed information about the region in an efficient way.

Currently, the Internet is the number one source for obtaining geographical information about interesting places worldwide. In this case, the Geographical Information Systems (GIS) embedded in the web services play a significant role in the distribution of the data towards the users. These users can be local people, but also customers, clients, businessman, tourist, etc. from across the globe. As a result, the demand for a geographical database which supports multiple languages (and writing scripts) is increasing rapidly. The web services thus have to be able to distribute, portray and process time-variant spatial data for multi-lingual stakeholders, visitors, investors and even decision makers from different language communities.

Some arguments have risen which suggest that, over time, the Internet would develop its own language: some simplified form of the English. This Internet-language would facilitate the global communication drastically, presuming that it also covers geographical names. However, at this moment there is no guarantee if and when this common language would be present. In any case, a link still has to be maintained between this Internet-language and the 'real' local geographical names (and the local writing scripts). The local signposts and road signs will not mention these Internet-names, which means that the people visiting the place need to be able to retrieve the original names.

In Xinjian, the research project about the Multilanguage Base Map Information System (MBMIS) for the main Digital Xinjian (DXJ) Project was initiated to meet the problems described above. Language support on web services is usually limited by the underlying technology: the operating system and application programs, such as the database management system. Certain GIS support the creating of web services and provide a wide range of customization options which enables the operators to generate intuitive and easy-to-use interfaces. (ESRI, 2009)

Con formato: Sin Resaltar

For the project described in this paper, Windows XP is selected as operating system. Furthermore, ESRI's software ArcGIS is considered to be the most suitable to embed the geographical information

in the web service. Both systems support most of the major world languages by Unicode character set as well as offering extensive functionality and allowing for full customization of a wide variety of parameters. As a pilot project, Uyghur, Chinese, English and Russian user interface versions and database are tested, whose results are presented in the next sections.

2. Method and Base Map Database

ArcGIS Desktop supports 136 locales and the following 17 language groups on Windows XP: Arabic, Greek, Simplified Chinese, Traditional Chinese, Armenian, Hebrew, Baltic, Indic, Turkish, Central Europe, Japanese, Vietnamese, Cyrillic, Korean, Western Europe, Georgian, Thai, and United States English. In the system, all these languages are supported with Unicode. (ESRI, 2009)

Con formato: Sin Resaltar

Unicode is a group of character encoding formats that supports most of the world's major languages. Several formats of Unicode are available today, including UTF-8, UTF-16 and UTF-32. It is becoming a popular encoding format as more data contains characters of multiple languages. (Unicode Inc, 2009) The format provides a unique number for every character, independent of the platform (operating system), the application program or the language which is used. Fundamentally, computers just deal with numbers, in a binary format, storing letters and other characters by linking a unique number to each one of them. The number in the format names mentioned above indicates the number of bits the format uses to encode a character. UTF-16, for example, uses 16 bits or 2 bytes for each character, resulting in 2^{16} or 65,536 possibilities. UTF-32 on the other hand makes it possible to encode over a million different characters.

Con formato: Sin Resaltar

Before Unicode was invented, there were hundreds of different encoding systems to assign these numbers. A single encoding can never contain all the characters. For example, the European Union alone requires several different encodings to cover all its languages. When only considering the English language, no single encoding is adequate to support all the letters, punctuation, and technical symbols commonly used. (Unicode Inc, 2009) These different encoding systems also conflict with one another: two encodings can use the same number for two different characters, or use different numbers for the same character. Any computer (especially servers) needs to support many different encoding systems; yet whenever data is passed between different encodings and/or platforms, misinterpretation or even corruption may occur. (UKIJ, 2009)

Con formato: Sin Resaltar

Con formato: Sin Resaltar

A brief analysis of the usage of Unicode shows that it is an indispensable element in most multi-lingual environments which are characterized by either the usage of several languages in the same document or the frequent exchange of data from different native languages. (Masumoto et al., 2005; N.N., 2002) If, on the other hand, the GIS is not ready to properly display and handle those other native languages, it will leave its user helpless. (Unicode Inc, 2009) As Uyghur is a member of the Turkish language group and its writing script originates from Arabic characters, Uyghur characters are supported in Windows XP, but not in the default case. Therefore, a special input

Con formato: Sin Resaltar

Con formato: Sin Resaltar

method and Uyghur Unicode fonts are needed. (Oyghan, 2009; UKIJ, 2009) The other three writing scripts are supported by default: Latin characters (English), (Simplified) Chinese characters and Cyrillic characters (Russian).

In the GIS domain, most users may not feel the need to use the Unicode system, even not on the Internet. But when different writing scripts are concerned, a level of caution is needed to avoid interpretation problems. However, once Unicode becomes a standard on the Internet – which would ameliorate the communication drastically – many (geographical) data providers will need to comply with it. Furthermore, a number of authors believe that the complications of including such standard to a certain software package such as ArcGIS could be cumbersome, if it is not taken into consideration at an early enough stage and prepared for. ArcGIS Desktop applications, such as ArcMap, are Unicode based, so they support Unicode to a certain level, but this level depends on the data format. (Unicode Inc, 2009)

Currently, a personal geodatabase is the only data format that supports Unicode by default. (ESRI, 2009) It is even possible to store and display characters of multiple languages in a single personal geodatabase which make applications with Uyghur characters possible in a GIS, but there are still some problems with the automation of the selection of the proper font. If the characters are not displayed correctly, it is necessary to verify that the font is set to Unicode, such as Tahoma or Microsoft Uighur. Since the ArcGIS geodatabase supports Unicode characters, all data stored in the geodatabase data model support a multi-lingual geodatabase. The 1:1 million scale base map data of Xinjiang Uyghur Autonomous Region was imported to the geodatabase and the multilingual information for the place name was added.

This project involves Multi-lingual information in Uyghur, Chinese, English and Russian. These diverse characters in Unicode are respectively distributed in Arabic, Simplify Chinese, Latin and Cyrillic character sets. Since only the Uyghur characters are written and read from right to left it is rather complicated to utilize it in combination with the other three character sets and Arabic numerals.

The recommended Uyghur input method, proposed by the Uyghur Computer Science Association "Oyghan Uyghur Unicode IME 3.0" on its website (http://www.ukij.org/oyghan/), is introduced to the system as a Uyghur input method for both Windows XP Professional version and ArcGIS 9.0. The Uyghur fonts used in the multilingual user interface for the desktop application, downloaded from http://www.ukij.org/oyghan/, are: UKIJ Basma, UKIJ Tuz, UKIJ Tuz Tom, UKIJ Esliye and Uyghur Tuz Unicode… and Microsoft Uighur in Windows Vista ™. (Oyghan, 2009; UKIJ, 2009)

3. The place names database in multilanguage

One main factor in the usability of the base map information is the diversity of languages, particularly in Xinjiang where a large number of (very diverse) languages and writing scripts are used beside each other. Consequently, the maps for general use would need to be multilingual: place names and map

Con formato: Sin Resaltar
Con formato: Sin Resaltar
Con formato: Sin Resaltar
Con formato: Sin Resaltar
Con formato: Sin Resaltar
Con formato: Sin Resaltar

legends ought to be presented in multiple languages though not necessarily all at once. Uyghur and Chinese are selected as the dominant languages while English and Russian are considered as international languages, since Russian is more commonly spoken in Central Asia than English. The problems with these place names include:

- lack of place name information about this rural area and its geographical location,
- misspelling and -pronunciation,
- variations in the names in English and in Chinese Pinyin caused by Chinese pronunciation.

In order to solve these problems fields for the other three languages, besides Uyghur, are added after importing the database of the original base map to the system geodatabase. Next, the place names are subsequently translated and checked in an initial test. To use Uyghur Latin Character (ULY) another extra field was added to the geodatabase. The structure of this geodatabase is presented in Table 1 and Table 2 depicts an extract of the attributes in the database.

The system facilitates the search of place names (in the different languages and writing scripts) on the Internet towards its users. Furthermore, to support the correct pronunciation of a place name in the local language, voice recordings in Uyghur and Chinese are included in a multimedia database linked with a link to the database with the place names. This link is also depicted in Table 3, last column. Not only the local oral communication benefits from this but it also supports the exact place name presentation to all non-native speakers.

Table 1 The structure of place name geodatabase

Field	Data Type	Description
ID	Integer	ID
Geo_ID	Integer	Geo_ID
Name_Cn	Unicode String	Name in Chinese
Name_Py	Unicode String	Name in Chinese Pinyin
Name_Ui	Unicode String	Name in Uyghur (Arabic)
Name_Ul	Unicode String	Name in Uyghur (Latin)
Name_En	Unicode String	Name in English
Name_Ru	Unicode String	Name in Russian
Pron_Ui	Hyper link	Pronunciation in Uyghur
Pron_Ch	Hyper link	Pronunciation in Chinese

Eliminado: 1
Con formato
Con formato
Con formato
Con formato
Con formato
Con formato
Con formato
Con formato
Con formato
Con formato
Con formato

Table 2 Attributes of the place name geodatabase

Eliminado: 2

Attributes of XJ_County93

PERIMETER	COUN	Uighur Name	Chinese Name	Pinyin	English Name	
800209	652222	باريكۆل	巴里坤哈萨	balikunhashake	Barikol	<Null>
542568	650121	ئۈرۈمچى ناھىيىسى	乌鲁木齐县	wulumuqixian	Urumqi County	<Null>
1050390	652827	خېجىڭ	和静	hejingxian	Hejing	<Null>
43497.800781	650101	ئۈرۈمچى	乌鲁木齐市	wulumuqishi	Urumqi	<Null>
587273	652922	ئونسۇ	温宿	wenshuxian	Onsu	<Null>
568310	652926	باي	拜城	baichengxian	Bay	<Null>
600066	652923	كۇچا	库车	kuchexian	Kucha	<Null>
429200	652927	ئۇچتۇرپان	乌什	wushenxian	Uchturpan	<Null>
582314	653023	ئاقچى	阿合奇	aheqixian	Akchi	<Null>
655541	652223	ئاراتۇرۇك	伊吾	yiwuxian	Aratuwruk	<Null>

Record: 0 Show: All Selected Records (0 out of 88 Selected.) Opti

Generally, the place names in Xinjian have two different pronunciations. For example, Urumqi (Urumchi) is also spelled in Chinese Pinyin as 'Wu u muqi'. In this example, the pronunciation is similar so it is not too difficult to distinguish it for local people or visitors, but problems start to rise when a user wants to retrieve information about it from in the Internet. More examples of these pronunciation differences are shown in Table 3. Currently, a method for correcting these sorts of problems in the future is under development.

Table 3 Examples of Pronouciation differences

Eliminado: 3

Name_Ch	Name_Py	Name_Ui	Name_Ul	Name_En	Pron_Ui	Pron_Ui
Name in Chinese	Pinyin in Chinese	Name in Uyghur	ULY (Uyghur Latin Character)	Name in English	Pronunciation in Uyghur	Pronunciation in Chinese
	Wu lu mu qi	ئۈرۈمچى	Urumchi	Urumqi	650111.mp3	650111_C.mp3
	Kashi		Qeshqer	Kashighar	653101.mp3	653101_C.mp3
	Shayibage		Saybagh	Saybagh	650102.mp3	650102_C.mp3
	Shache		Yerken	Yarkan	653122.mp3	653122_C.mp3
	Ruoqiang		Charqiliq	Qarkilik	652827.mp3	652827_C.mp3
	Aletai		Altay	Altai	654301.mp3	654301_C.mp3

4. The Multilingual Graphic User Interface and Output

Currently, the Multilingual GUI of ArcGIS desktop and the web client of ArcIMS are in a testing phase to make it more stable and flexible. (ESRI, 2009) Data frame and layer names are successfully displayed in all characters in the same view. Figure 1 depicts such a test-example where the layer names and labels are displayed in the same view at the same time in Uyghur, Chinese, Russian and English in different layers and combined in the same layer in ArcGIS Desktop.

Con formato: Sin Resaltar

The legend automation, automatic labeling and annotation also worked fine within ArcGIS desktop (see Figure 3), but some problems occurred when testing it in ArcIMS. However, the names of the layers in non-roman scripts characters are displayed as a question mark (See Figure 2). Tool tips and menus in the Uyghur language are not tested yet.

Since Geodatabase support Unicode, it is possible to implement the labeling in four languages through the Maplex labeling extension of ArcGIS which allows cartographic output in digital or printed form. After several tests, it was concluded that the Uyghur annotation in ArcGIS Desktop does not require Unicode. However as mentioned before, it is better to use Unicode to support multiplatform use.

Figure 4 shows the results of the tests concerning queries for the place names in the different languages and writing scripts on the multi-lingual database. Both in ArcGIS Desktop and in ArcIMS, these result were successful. Figure 5 depicts an extract from the multi-lingual database.

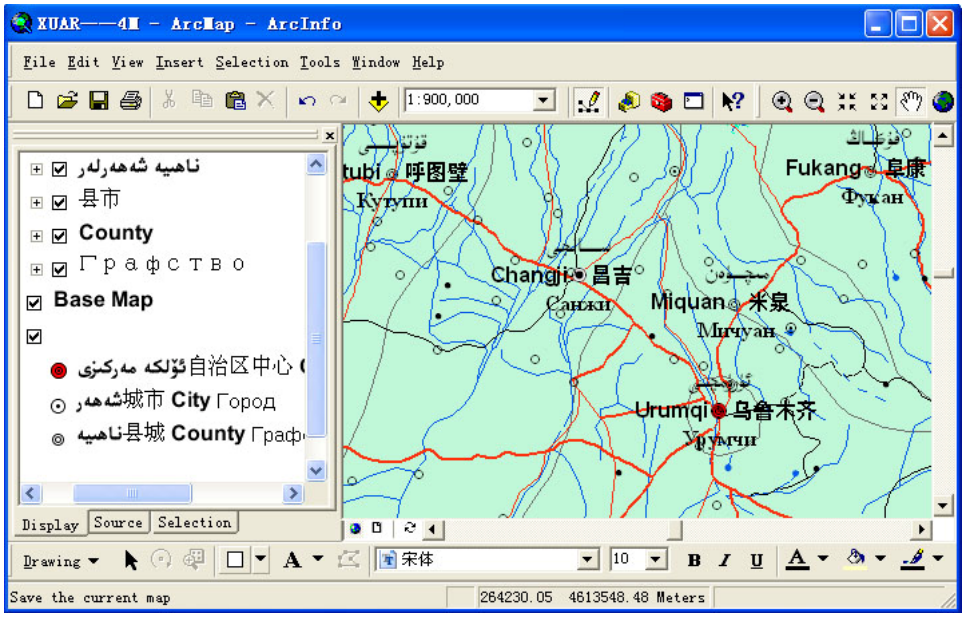

Figure 1 Display of the layer names and labels in ArcGIS Desktop

Eliminado: 1

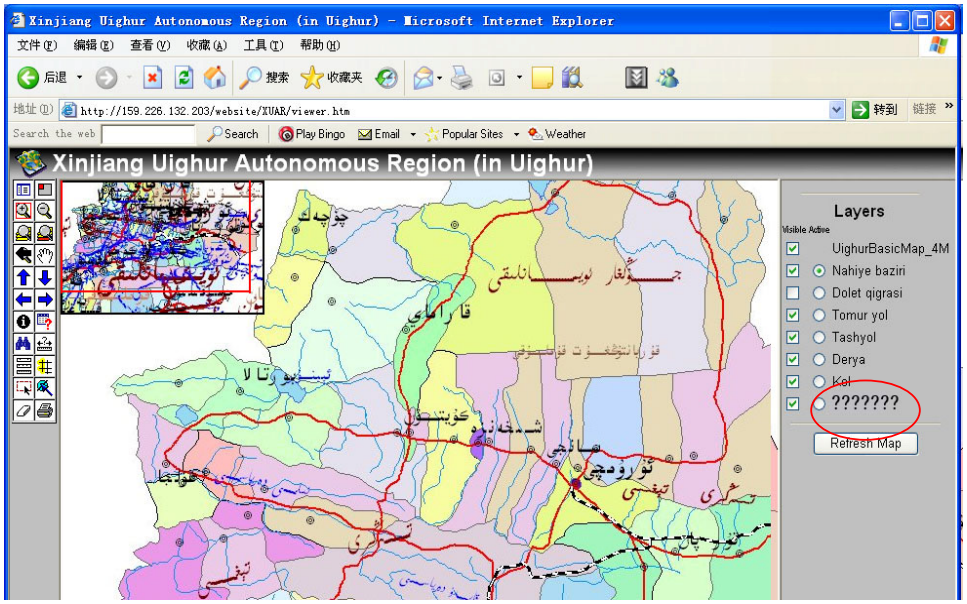

Figure 2 Problems with the display labels in a non-roman script.

Eliminado: 2

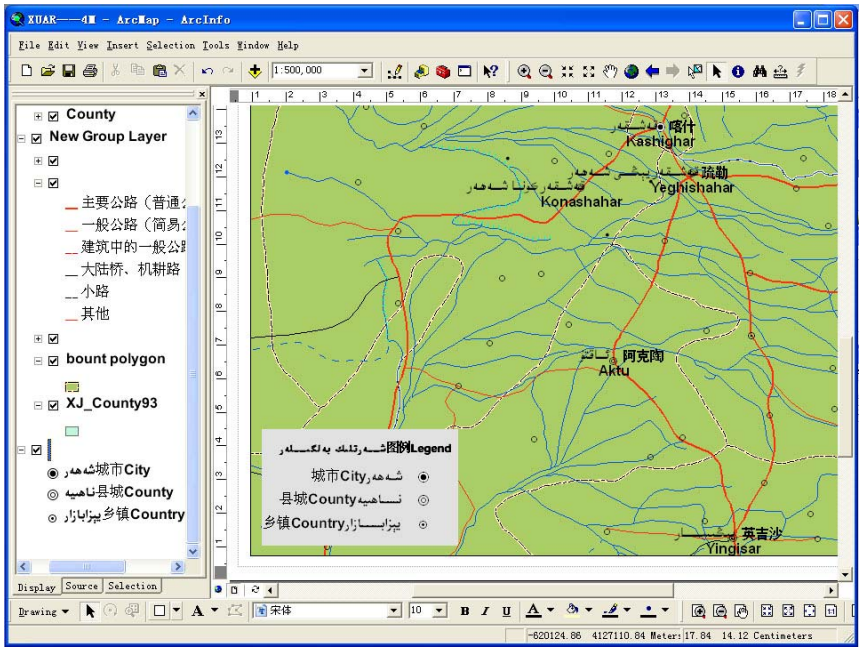

Figure 3 Legends are displayed correctly at the same time automatically in ArcGIS desktop.

Eliminado: 3

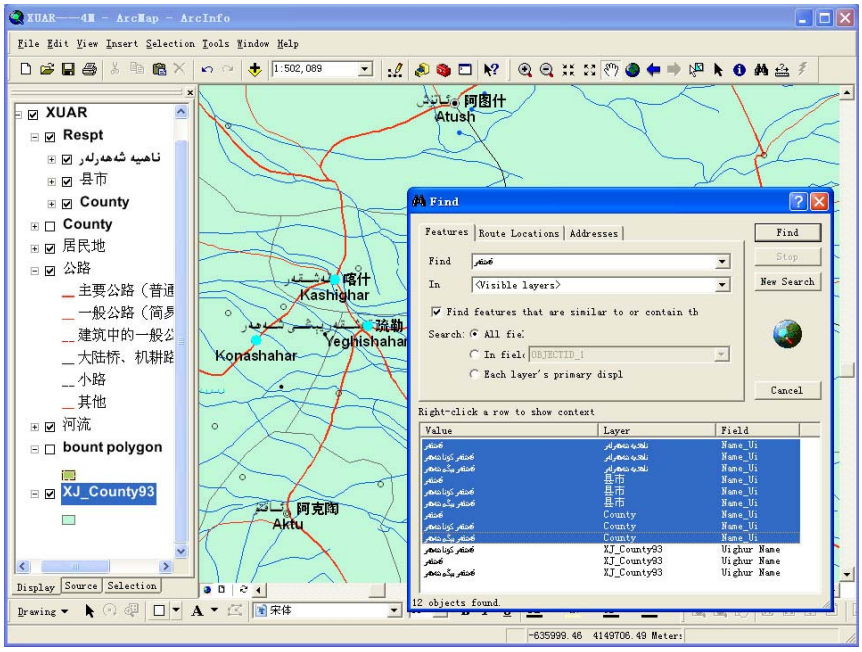

Figure 4 Multi-lingual querying by attribute data are successful both in ArcGIS desktop and ArcIMS.

Eliminado: 4

Figure 5 The Uyghur character fields can be used as acommon field for making relation and links between tables.

Eliminado: 5

5. Problems

ArcGIS still has problems with the display of the Uyghur Characters in automatic labeling, annotation and legend. The Spline text function is often used to allow text to be placed along a (curved) line. However, this does not work well for the Uyghur characters: they become unreadable since all the characters and words are separated and spread along the line. This strategy can be applied to the other writing scripts – such as the Latin script –, but not to the Arabic script as the subsequent characters need to preserve a link with each other. This separation of Uyghur characters along a Spline is depicted in Figure 6.

Furthermore, when the labeling function and the annotation are used, all the Uyghur characters will get longer, as well as, the space between words. The other three character types do not show this problem.

In Figure 7 the stretching of the characters and word spaces is illustrated. The table on the left shows the correct writing; the illustrations on the right are extracts from the map. The horizontal lines in these characters are much longer than they should be.

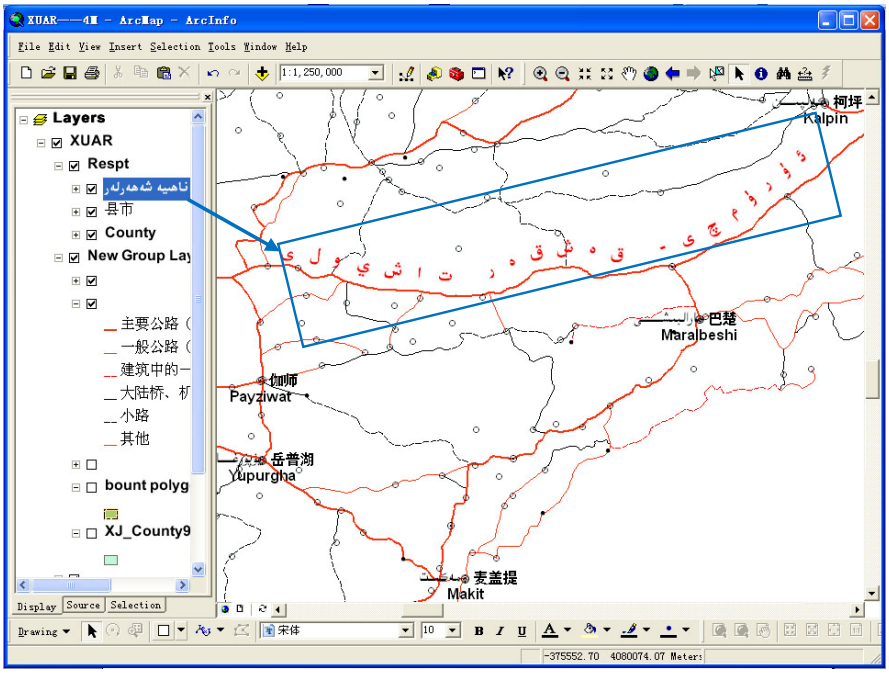

Figure 6 Problems in the ArcGIS Desktop interface with the Uyghur characters: character separation along a Spline

Eliminado: 6

Uighur Name
قۇبۇقسار
چاغانتوقاي
تولى
چىڭگىل
قاراماي
ئارشاڭ

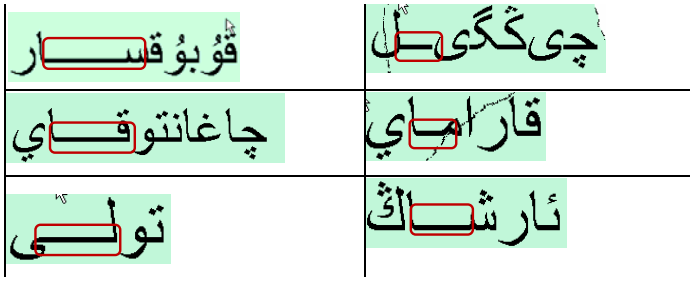

Figure 7 Problems in the ArcGIS Desktop interface with the Uyghur characters: length of characters

6. Conclusion

During the tests, several problems, mainly related with the correct display of the – from Arabic originating – Uyghur characters, where discovered. On the other hand, it became clear that the Unicode system is improving rapidly. Although there are still some disadvantages, such as the double size of data compared to the normal size of normal code pages, Unicode promises to be the best method since data storage becomes less of an issues. The correct display for Uyghur characters is however still the main struggle point in the web application presented here. Further tests are planned to overcome the problem described above.

References

UKIJ (2009)Uyghur Kompyutér Ilimi Jem'iyiti . [online] Available at: www.ukij.org

Oyghan (2009) Uyghur Unicode fonts for Microsoft Windows. [online] Available at: www.oyghan.com

Unicode, Inc (2009) Unicode Home Page. [online] Available at: http://www.Unicode.org

ESRI (2009) ArcIMS, Frequently Asked Questions. [online] Available at: http://www.esri.com/software/arcgis/arcims/faq.html

ESRI (2009) ArcGIS 9.2 Desktop Help. [online] Available at: http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=welcome

Xiaoyan J. (2006) Discussions on the Multi-source Data Processing Technology in the Construction of GIS Databases, GSDI-9 Conference Proceedings, Santiago, Chile

Masumoto S., Raghavan V., Nonogaki S., Neteler M., Nemoto T., Mori T., Niwa M., Hagiwara A. and Hattori N.(2005) Multi-Language Support and Localization of GRASS GIS, International Journal of Geoinformatics. 1:1.

Deckert C. (2004) Modeling Language Diffusion with ArcGIS, ESRI International User Conference Proceedings. San Diego, CA, pp 25.

N.N. (2002) Proceedings of the Workshop, Asian Language Resources and International Standardization, Center of Academia Activities, Academia Sinica, Taipei, Taiwan

Veselionova L. and Booza J. (2006) Using GIS Map to the Multilingual City, ESRI International User Conference Proceedings. San Diego, CA, pp 13.

ISO-19115; Geographic information -- Metadata