

Date: 20/07/2006

#### Unicode turn, BULAC experience

#### **Bernard Desnoues**

Bibliothèque universitaire des langues et civilisations Paris, France

Meeting:	77 UNIMARC
Simultaneous Interpretation:	Yes

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL

20-24 August 2006, Seoul, Korea http://www.ifla.org/IV/ifla72/index.htm

#### Abstract:

BULAC, the Bibliothèque universitaire des langues et civilisations (in English: University Library for Language and Civilization Studies) will open to the public in 2010. It will offer a wide range of documentary resources on Non-Western languages and civilizations to students and researchers. These resources are at present scattered in many locations and are to be gathered in the new library.

The primary aim of the project is to build a common multi-script catalogue. We had first of all to convert data from previous computerized catalogues (several formats, several character sets). We purchased consequently an integrated library system in 2003. This library system was decided to be unimarc and unicode compliant according to our specifications.

The catalogue was launched in july 2005. It contains by now CJK, Cyrillic, Greek and Latin scripts. But, we had many difficulties with the character sets conversion. We had also to understand Unicode principles.

The catalogue is now working properly. But we have to improve the index rules and we have also to implement many other scripts in the catalogue.

#### Diapositive 1



# Unicode turn BULAC experience

by Bernard DESNOUES

Librarian

University Library for Language and Civilization Studies - Paris

#### Diapositive 2



### Introduction (1)

BULAC, the University Library for Language and Civilization Studies



http://www.bulac.fr

Under the supervision of the French department of Higher Education and Research, the University Library for Language and Civilization Studies was created to form the core of a European-level centre of excellence.

Bulac will gather document collections coming from more than 20 libraries all located in Paris:

- $\sim 1500000$  documents
- $\sim 25~000$  serials



### Introduction (2)

BULAC will be devoted to Non-Western languages and civilizations



## Central and eastern Europe, Asia, Africa, Oceania and the Amerindian civilizations

One half of the 1.5 million documents are written in Non-Latin scripts. Many languages and scripts are present in our collections :

- more than 100 languages
- more than 40 scripts

#### Diapositive 4



### Introduction (3)

## BULAC will open to the public in 2010



## **BULAC** will be situated very near the Bibliothèque nationale de France

BULAC will open in a new building whose construction will start in 2007 in a new fledgeling district gathering numerous university libraries, by the Bibliothèque nationale de France.



### The aims of the project





### Bulac catalogue (1)

- to build a catalogue was the first important realization of BULAC http://catalogue.bulac.fr
- launched in july 2005
- primarily made from heterogeneous data
  - several formats (Unimarc, Marc 21)
  - several character sets (ISO 5426, ANSEL, BIG5, Shift-JIS, KSC, EACC)
- to unify data before unifying the collections

This action had a high priority because some libraries which are taking part in the project had no computerized catalogue. BULAC purchased consequently an integrated library system in 2003 to build a common catalogue.



### Bulac catalogue (2)



Our integrated library system is Millennium, a product from Innovative interfaces.



### Bulac catalogue (3)

- BULAC is part of the French union catalogue: the Sudoc (Système universitaire de documentation) http://www.sudoc.abes.fr
- BULAC catalogue is daily updated by the Sudoc
- current cataloguing but also important retrospective conversion of older card catalogues

The French Sudoc turned to Unicode recently (june 2005). So we can now catalogue documents written in CJK, Cyrillic, Greek, Armenian, Georgian. The right to left scripts are not yet supported. An other restriction: we cannot till now load unicode data from Sudoc in our library system and consequently we only receive the Latin part of the records in our catalogue.

#### Diapositive 9



### Bulac catalogue (4)

- statistics
  - more than 500 000 bibliographic records
  - more than 580 000 item records
  - growing 500/600 bibliographic records per a day

The daily increase of the catalogue is largely due to the retrospective conversion of cards catalogues. When these actions will be over, the catalogue will contain more than 1 500 000 bibliographic records.



### BULAC catalogue principles

- a working group defined the principles to be applied to the BULAC catalogue in a preliminary report (jan. 2002) :
  - Unicode multi-script cataloguing
  - marc format
  - Unicode compliant library system
  - transliteration standards allowing easy conversions
  - multi-script authorities

A working group met 7 times between may and december 2001 to study the Non-Latin collections in French libraries and defined the above principles. A report has been written: http://www.sup.adc.education.fr/bib/Acti/fcnl/sommaire.htm, in french.



### A multi-script catalogue?

- working : Arabic, Chinese, Cyrillic, Japanese, Korean, Persian and of course Latin
- testing : some Indian scripts (Devanagari, Malayalam), Lao
- in the future: many other scripts such as Thai, Tibetan, and some African scripts



### Why did we choose Unimarc (1)

#### multi-script compliant

- ■200 1#|6a02|7ea|a1992 現代陶藝國際邀請展圖錄|d= The 1992 International Invitational Exhibition of Contemporary Ceramic Art|f[主編高玉珍]
- 200 1#|6a02|7ba|a1992 xian dai tao yi guo ji yao qing zhan tu lu|bTexte imprimé|d= The 1992 International Invitational Exhibition of Contemporary Ceramic Art|f[zhu bian Gao Yuzhen]
- the cataloguing format of the French union catalogue : Sudoc

The Unimarc manual, section 3, § 3.9 and 3.10 defines the treatment of different scripts.



### Why did we choose Unimarc (2)

#### Unicode compliant

- ■200 1#|6a02|7ea|a1992 現代陶藝國際邀請展圖錄|d= The 1992 International Invitational Exhibition of Contemporary Ceramic Art|f[主編高玉珍]
- 200 1#|6a02|7ba|a1992 xian dai tao yi guo ji yao qing zhan tu lu|bTexte imprimé|d= The 1992 International Invitational Exhibition of Contemporary Ceramic Art|f[zhu bian Gao Yuzhen]

To show that Unimarc is Unicode compliant is very difficult! Unimarc and Unicode are in fact independant things. Unimarc is simply a structured container for data.



### Screenshots (1)

#### cataloguing: Arabic script

الحمد f تطور الفكر السياسي الشيعي من الشورى الى ولاية الفقيه 6a02|7fa|a/المدور الكاتب

2001#]6a02|7ba|aTaṭawwur al-fikr al-siyāsī al-šīʿī min al-šūrá ilá wilāyat al-faqīh|bTexte imprimé|fAḥmad al-Kātib 210##|6a04|7fa|a: ועניסן אַנעיסן, |d1998

210##|6a04|7ba|aBayrūt|cDār al-Jadīd|d1998



### Screenshots (2)

#### opac: Arabic script

AUTHOR Kātib, Aḥmad al-

کاتب, احمد

Tatawwur al-fikr al-siyāsī al-šī'ī min al-šūrá ilá wilāyat al-faqīh [Texte imprimé] / Aḥmad al-Kātib

تطور الفكر السياسي الشيعي من الشوري الى ولاية الفقيه / احمد الكاتب

EDITION al-Tab'ah al-Bayrūtīyah 1

الطبعة البيروتية 1

IMPRINT Bayrūt: Dār al-Jadīd, 1998

بيروت : دار الجديد, 1998



### Screenshots (3)

#### cataloguing: Arabic script (Persian)

تطور الفكر السياسي الشيعي من الشورى الى ولاية الفقيهa|7fa|2001#|6a02|7fa| احمد الكاتباً|

2001#|6a02|7ba|aTaṭawwur al-fikr al-siyāsī al-šī'ī min al-šūrá ilá wilāyat al-faqīh|bTexte imprimé|fAḥmad al-Kātib

| d1998ماردار الجديدي| بيروتa|7fa|a|#|6a04|7fa|

210##|6a04|7ba|aBayrūt|cDār al-Jadīd|d1998



### Screenshots (4)

#### opac : Arabic script (Persian)

AUTHOR Kātib, Ahmad al-

کاتب, احمد

Taṭawwur al-fikr al-siyāsī al-šī'ī min al-šūrá ilá wilāyat al-faqīh [Texte imprimé] / Aḥmad al-Kātib

تطور الفكر السياسي الشيعي من الشورى الى ولاية الفقيه / احمد الكاتب

EDITION al-Ţab'ah al-Bayrūtīyah 1

الطبعة البيروتية 1

IMPRINT Bayrūt: Dār al-Jadīd, 1998

بيروت: دار الجديد, 1998



### Screenshots (5)

#### cataloguing : Cyrillic script

200 1#|6a01|7ca|aБолгарский хронограф и его судьба на Руси|bTexte imprimé|fЛюдмила В. Горина|g пред. Милияна Каймакамова

200 1#|6a01|7ba|aBolgarskij hronograf i ego suďba na Rusi|bTexte imprimé|fLûdmila V. Gorina|g[pred. Miliâna Kajmakamova]

700 #|6a05|7ca|3102369496|аГорина|bЛюдмила Василиевна|f1933-....|4070

700 #|6a05|7ba|3102369496|aGorina|bLûdmila Vasilevna|f1933-....|4070



### Screenshots (6)

#### opac: Cyrillic script

AUTHOR Горина, Людмила Василиевна (1933-....) Auteur. Gorina, Lûdmila Vasilevna (1933-....) Auteur.

ПТLE Болгарский хронограф и его судьба на Руси [Texte imprimé] / Людмила В. Горина; пред. Милияна Каймакамова Bolgarskij hronograf i ego sud'ba na Rusi [Texte imprimé] / Lûdmila V. Gorina; [pred. Miliâna Kajmakamova]

IMPRINT София: ТАНГРА ТанНакРа, 2005

Sofiâ : TANGRA TanNakRa, 2005



### Screenshots (7)

#### cataloguing: Greek script

- 200 1#|6a01|7ga|aΣεφέρης|bTexte imprimé|eμε τους ρύθμους των άλογων|fΜιχάλης Τσιανίκας
- 200 1#|6a01|7ba|aSeférīs|bTexte imprimé|eme tous rýthmous tōn álogōn|fMichalīs Tsianíkas
- 210 ##|6a02|7ga|a[Αθήνα]|cΕκδόσεις Κανάκη|dcop. 2000
- 210 ##|6a02|7ba|a[Athīna]|cEkdóseis Kanákī|dcop. 2000
- 700 #1|6a04|7ga|3092877974|aΤσιανίκας|bΜιχάλης|4070
- 700 #1|6a04|7ba|3092877974|aTsianíkas|bMichális|4070



### Screenshots (8)

#### opac: Greek script

**ΑυτΗΟΚ** Τσιανκα, Μιχλη Auteur.

Tsianíkas, Michális Auteur.

TITLE Σεφρη [Texte imprimé] : με του ρθμου των λογων / Μιχλη Τσιανκα Seférīs [Texte imprimé] : me tous rýthmous tôn álogôn / Michalīs

Tsianíkas

ΙΜΡΡΙΝΤ [Αθνα] : Εκδσει Κανκη, cop. 2000

[Athīna]: Ekdóseis Kanákī, cop. 2000



### Screenshots (9)

#### cataloguing: Korean script

200 1#|6a01|7ka|a다시 쓴 우리말 어원이야기|bTexte imprimé|f조항범

200 1#|6a01|7ba|aTashi ssŭn uri mal ŏwŏn iyagi|bTexte imprimé|fCho Hang-bŏm

210 ##|6a01|7ka|a서울|c한국문원|d1997

210 ##|6a01|7ba|aSéoul|cHan'guk munwŏn|d1997

700 #1|6a01|7ka|a조|b항범|f1958-...|4070

700 #1|6a01|7ba|aCho|bHang-bŏm|f1958-....|4070



### Screenshots (10)

#### opac: Korean script

AUTHOR 조, 항범 (1958-...)

Cho, Hang-bom (1958-....) Auteur.

TITLE 다시 쓴 우리말 어원이야기 [Texte imprimé] / 조항범

Tashi ssŭn uri mal ŏwŏn iyagi [Texte imprimé] / Cho Hang-bŏm

IMPRINT 서울: 한국문원, 1997

Séoul: Han'guk munwŏn, 1997



### Screenshots (11)

#### cataloguing : Chinese script

- 200 1#|6a01|7ea|a海上人家|bTexte imprimé|e海洋漁業經濟與漁民社會|f歐陽宗書著
- 200 1#|6a01|7ba|aHaishang renjia|bTexte imprimé|eHaiyang yuye jingji yu yumin shehui|fOuyang Zongshu zhu
- 210 ##|6a01|7ea|a南昌|c江西高效出版社|d1998
- 210 ##|6a01|7ba|aNanchang|cJiangxi gao xiao chu ban she|d1998
- 700 #1|6a01|7ea|a歐陽|b宗書|4070
- 700 #1|6a01|7ba|aOuyang|bZongshu|4070



### Screenshots (12)

#### opac: Chinese script

AUTHOR 歐陽, 宗書 Auteur.

Ouyang, Zongshu Auteur.

TITLE 海上人家 [Texte imprimé]:海洋漁業經濟與漁民社會 / 歐陽宗書著 Haishang renjia [Texte imprimé]: Haiyang yuye jingji yu yumin shehui / Ouyang Zongshu zhu

IMPRINT 南昌: 江西高效出版社, 1998

Nanchang: Jiangxi gao xiao chu ban she, 1998



### Screenshots (13)

#### cataloguing : Japanese script

200 1#|6a01|7da|a図解 考古学辞典|bTexte imprimé|f水野清一、小林行雄

200 1#|6a01|7ba|aZukai Kōkogaku jiten|bTexte imprimé|fMizuno Seiichi, Kobayashi Yukio

210 ##|6a01|7da|a東京|c倉元新社|d1964

210 ##|6a01|7ba|aTōkyō|cSōgen shinsha|d1964

700 #1|6a01|7da|a水野|b清一|f1905-1971|4070

700 #1|6a01|7ba|aMizuno|bSeiichi|f1905-1971|4070

701 #1|6a01|7da|a小林|b行雄|f1911-1989|4070

701 #1|6a01|7ba|aKobayashi|bYukio|f1911-1989|4070



### Screenshots (14)

#### opac: Japanese script

AUTHOR 水野, 清一 (1905-1971) Auteur.

Mizuno, Seiichi (1905-1971) Auteur.

TITLE 図解考古学辞典 [Texte imprimé] / 水野清一、小林行雄

Zukai Kōkogaku jiten [Texte imprimé] / Mizuno Seiichi, Kobayashi Yukio

IMPRINT 東京: 倉元新社, 1964

Tőkyő: Sőgen shinsha, 1964



### Multi-script authorities

#### searching Non-Latin forms

■200 #1|aZola|bÉmile|f1840-1902 300 ##|aÉcrivain français 400 #0|5e|aAgrippa 400 #1|7ba0yda0y|8frejpn|aゾラ|bエミール



### Remarks about Unimarc (1)

- Unimarc script code list
  - the script code list is too limited: 15 codes are at present defined
  - the "zz" code is used for all remaining names of scripts
- why not adopting the ISO 15924 standard (4 characters codes)?
  - very complete script code list
  - the Unicode consortium was appointed as the registration authority for this standard and use it as script name reference
  - more precise statistics
  - but this would require changes in the format

But we might wonder if the script codes are strictly necessary if the script coding system is Unicode because each character has its code and the display should always be good. The script code is useful to indicate the used script for non-specialists. It may also be used to allow some automatized treatments (like automatized romanization of some scripts like Cyrillic, Greek, ...). It could perhaps be used for the right to left scripts... to make the software align these fields on the right of the screen.



### Remarks about Unimarc (2)

•we chose to systematically duplicate non-Latin fields with transliterated Latin fields

200 1#|6a01|7jpan|a中 世 東 国 の 鉄 文 化 解 明 の 前 提|bTexte imprimé|f福 田 豊 彦

200 1#|6a01|7latn|aChūsei Tōkoku no tetsubunka kaimei no zentei|bTexte imprimé|fFukuda Toyohiko

- each alternative field contains a subfield \$7 containing a field script code
  - even if the script of the field is coded in the 100 field



### Remarks about Unimarc (3)

#### NSB / NSE

- these special codes primarily defined in the ISO 6630 standard are not defined in Unicode and each library system provider has to define its own values
- for example in our catalogue : uE136 and uE137 (Unicode private area)
- trouble with right to left scripts
- trouble with data exchange?



### A soft revolution, Unicode

- Unicode is an international consortium
- BULAC is an associate member of it
- Unicode became an international standard: ISO 10646
- first release in 1991, current release 4.1
- more than 100 000 characters

You can get many informations through the internet site of the consortium: http://www.unicode.org. For French librarians, an excellent canadian site provides French translations: http://hapax.qc.ca and other pieces of information.



### Why did we choose unicode?

- libraries catalogues will contain more and more records containing alternate scripts
- we chose unicode because it simplifies the management of varied scripts
- but Unicode is a complex world...



### Some Unicode principles

- one character, one code
  - no need use escape sequences to control several scripts in records

880 1 \$6100-01/**(3/r**\$aEHf CHj GdEUHY, YHO GdYXje Hf YHO GdhGMO,\$d1198-1256

- doesn't deal with the display of characters
- compatibility code to code with older character sets

The compatibility is the reason why it may happen that the same character can be input in two ways (specially Latin characters with diacritics).



### First step, Unicode conversion

- from ISO 5426 and ANSEL
- from data containing two character sets
  - ISO 8859-1 and an other character set (Shift-JIS or Big5)
- from EACC

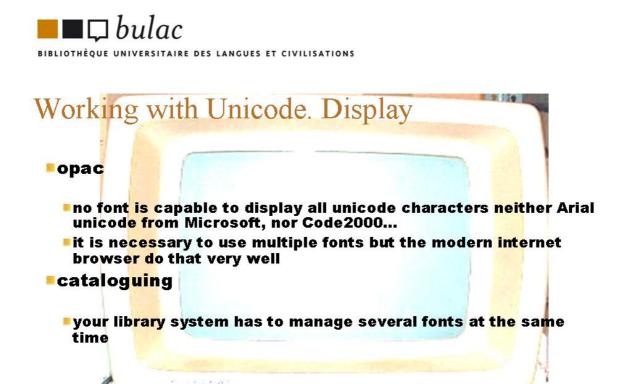
There were difficulties in ISO 5426 and ANSEL conversions because the diacritics comes before the characters which they apply to whereas the diacritics follow the characters according to Unicode. So the process has to identify the complete relevant sequences before conversion. A supplementary difficulty is the fact that the character is sometimes followed by more than one diacritic (see vietnamese language). For example :

ISO 5426 :  $\sim + ^ + ^ + o$  to convert in Unicode precomposed character :  $\tilde{o}$  (or canonical decomposition  $o + ^ + ^ + ^ - )$  which have to be equivalent forms for indexing.

When two character sets in the same file, we had to contend with codes conflicts between ISO 8859-1 and Shift-JIS or Big5 codes in some files. So some latin characters were badly converted. For example:

300 ##Trad. de : "Nankin sen : tozasareta kioku o tazunete : moto heishi 102 nin no shen", Tôoyô, Shakai heiron shoppansha, 2002.

We had to correct the data manually.

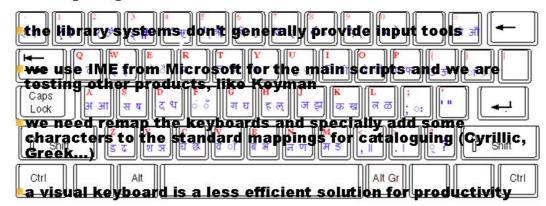


We install a set of fonts on each computer so that the opac can display all scripts we use and some we don't use at this time but we will use in the future. Our library system doesn't deal with several fonts. We must use one font at a time and it is very difficult to change that. But most cataloguers use only one script.



### Working with Unicode. Input tools (1)

#### cataloguing





### Working with Unicode. Input tools (2)

#### opac

- a catalogue for europeans or a universal catalogue ?
- what tools to search the catalogue with non latin scripts?
- visual keyboards may be a good way but not very useful for the public

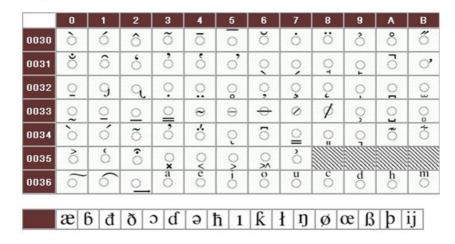
The search in the catalogue need the use of IME keyboards or other tools to input Non-Latin characters. This is not very useful for the public and specially for the remote public. That's why we always create transliterated fields in Non-Latin records so that the documents can be searched with a Latin keyboard.

Our public is mainly a European public and need searching with Latin keyboard.



### Working with Unicode. Input tools (3)

#### our project of visual tool to input diacritics and some other special characters



This experimental visual keyboard was built with Javascript. It might allow the public to input special diacritics and characters.



### Unicode trouble... (1)

- mixing right to left and left to right scripts
  - mirror characters caused trouble but the problem is at present solved

ADRESSE 1383 ,غيران : هيرمند , 2004]

Tihrān: Hirmand, 1383 [2004]

The display is not correct. The date into square brackets is rejected at the beginning of the field in the Arabic field. The cause is the "weak" directionality property of "mirror characters" such parenthesis, square brackets... There were difficulties with the Unicode bidirectional algorithm (alias Bidi algorithm, http://www.unicode.org/reports/tr9/). "This document describes the algorithm used to determine the directionality for bidirectional Unicode text. The algorithm extends the implicit model currently employed by a number of existing implementations and adds explicit format codes for special circumstances. In most cases, there is no need to include additional information with the text to obtain correct display ordering."

A library system provider has to implement this algorithm to manage properly the display of bidirectional data.



### Unicode trouble... (2)

#### index rules

- trouble when precomposed and canonical decomposition forms representing the same character are mixed in the database
- importance of normalization rule for data storage
- a library system is full unicode compliant if a normalization rule is applied
- Unicode provides a basic collation algorithm which can be modified as needed

Data contains both canonical decomposition and precomposed characters. A Unicode normalization rule is not applied. These equivalent forms are to be indexed in the same way. But it's not always true in our catalogue.

When implementations keep strings in a normalized form, you can be assured that alternative forms will have equivalent representations. See the Unicode reference document "Unicode normalization forms", http://www.unicode.org/reports/tr15/.

The Unicode consortium provides a specification for how to compare two Unicode strings while remaining conformant to the requirements of The Unicode Standard. See the "Unicode collation algorithm", http://www.unicode.org/reports/tr10/.



### Conclusion

- We came a long way
- But a lot of work is still to be done
  - implementing new scripts
  - improving indexation
  - make the input tools easier to use
- Unicode is a no return step
- Unimarc future ?



#### Contacts

Bibliothèque universitaire des langues et civilisations 60 rue de Wattignies

75012 - PARIS Tel: +33 (0)1 53 46 15 60 Fax: +33 (0)1 53 46 15 90

Director : Marie-Lise Tsagouria marie-lise.tsagouria@bulac.sorbonne.fr

Chief, cataloguing : Bernard Desnoues bernard.desnoues@bulac.sorbonne.fr

Chief, computing : Ginette Beurton ginette.beurton@bulac.sorbonne.fr