Unicode turn, BULAC experience

Bernard Desnoyes
Bibliothèque universitaire des langues et civilisations
Paris, France

Meeting: 77 UNIMARC
Simultaneous Interpretation: Yes

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.
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:
- ~ 1 500 000 documents
- ~ 25 000 serials
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
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

- to gather research teams and documentary resources
- to gather and harmonize collections scattered at present in many locations in Paris
- make documents access easier for French and foreign researchers and students
- make important and valuable documentary resources more visible
Bulac catalogue (1)

- to build a catalogue was the first important realization of BULAC
- launched in July 2005
- primarily made from heterogeneous data
  - several formats (Unimarc, Marc 21)
  - several character sets (ISO 5426, ANSI, 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.
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](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.
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 现代陶瓷国际邀请展图录|d= The 1992 International Invitational Exhibition of Contemporary Ceramic Art|f[主编高玉珍]
  - 200 1#|6a02|7ba|a1992 xian dai tao yi guo ji yao qing zhan tu lu|b Texte 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 现代陶艺国际邀请展图錄|d= The 1992 International Invitational Exhibition of Contemporary Ceramic Art|f[主编高玉珍]
  - 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 independent things. Unimarc is simply a structured container for data.
Screenshots (1)

**cataloguing : Arabic script**

أحمد | تطور الفكر السياسي الشيعي من الفئري الى ولاية الفقيه | الكاتب
2001 | aTaṭawwur al-fikr al-siyāsī al-šī‘ī min al-šūrā ilā wilāyat al-faqīh | bTexte imprimé | fAhmad al-Kātib
Screenshots (2)

**opac : Arabic script**

**AUTHOR** Kāṭib, Ahmad al-
کاتب، احمد

**TITLE** Tatawwur al-fikr al-siyāsī al-šī‘ī min al-šūrā ilá wilāyat al-faqīh
[Texte imprimé] / Ahmad al-Kāṭib
تطور الفكر السياسي الشيعي من الشورى إلى ولاية الفقيه / احمد الكاتب

**EDITION** al-Tab‘ah al-Bayrūtiyah 1
الطبعة بروتية 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 الفقيه

علي مسعود

2001\#\6a02\7fa\a

المشاركة في اجتماع الاوقاف

2001 (السعودية)

210\#\6a04\7fa\a

مكتبة الاوقاف

2001 (السعودية)

210\#\6a04\7fa\a

Dahir al-Jaddid

2001 (السعودية)
Screenshots (4)

**opac : Arabic script (Persian)**

**AUTHOR** Kātib, Aḥmad al-

**TITLE** Taṭawwur al-fikr al-siyāṣī al-šīʿī min al-šūrā ilā wilāyat al-faqīh

[Texte imprimé] / Aḥmad al-Kātib

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

**EDITION** al-Tab‘ah al-Bayrūṭīyah 1

الطبعة البتروتية 1

**IMPRINT** Bayrūt : Dār al-Jadīd, 1998

بيروت : دار الجديد, 1998
Screenshots (5)

cataloguing : Cyrillic script

200 1#6a017ca|aБолгарский хронограф и его судьба на Руси|bTexta
imprimé|fЛюдмила В. Горина|gпред. Миллиана Каймакамова
200 1#6a017ba|aBolgarski kjronograf i ego sud’ba na Rusi|bTexta
imprimé|fЛюдмила V. Gorina|g[pred. Milliana Kajmakamova]
700 #6a057ca|3102369496|aГорина|bЛюдмила Василиевна|f1933-
....|4070
700 #6a057ba|3102369496|aGorina|bLudmila Vasilevna|f1933--|4070
Screenshots (6)

- opac : Cyrillic script

**AUTHOR** Горина, Людмила Василяевна (1933–….) Auteur.
Gorina, Lûdmiła Vasilevna (1933–…) Auteur.

**TITLE** Болгарский хронограф и его судьба на Руси [Texte imprimé] / Лудмила В. Горина ; пред. Мильана Каемакамова
Bolgarski hronograf i ego sud’ba na Rusi [Texte imprimé] / Lûdmila V. Gorina ; [pred. Miliâna Kajmakamova]

**IMPRINT** София : ТАНГРА TanNakRa, 2005
Sofiá : TANGRA TanNakRa, 2005
Screenshots (7)

cataloguing: Greek script

200 1#|6a01|7ga|αΣεφέρης|bTeXte imprimé|ame tous rhythmus tôn αλογων|fMichális Tsianíkas
200 1#|6a01|7ba|αΣεφέρης|bTeXte imprimé|ame tous rhythmus tôn αλογων|fMichális Tsianíkas
210 ##|6a02|7ga|Αθήνα|cΕκδόσεις Κανάκη|dcop. 2000
210 ##|6a02|7ba|Αθήνα|cΕκδόσεις Κανάκη|dcop. 2000
700 #1|6a04|7ga|3092877974|aΤσιανίκας|bMichális|4070
700 #1|6a04|7ba|3092877974|aΤσιανίκας|bMichális|4070
Screenshots (8)

**opac : Greek script**

**AUTHOR** Tsianka, Michalis Auteur.
Tsianikas, Michalis Auteur.

**TITLE** Σφηρη [Texte imprimé] : με του ρθμου των λογων / Michalis Tsianka
Sefēris [Texte imprimé] : me tous rythmous tôn álogon / Michalis
Tsianikas

**IMPRINT** [Αθήνα] : Εκδοσι Kανκη, cop. 2000
Screenshots (9)

- **cataloguing : Korean script**

  200 1#|6a01|7ka|a 다시 쓴 우리말 어원이야기|b Texte imprimé|f 조항범
  200 1#|6a01|7ba|a Tashi ssŭn uri mal ᄀᆞᆫ ᄀᆞᆫ iyagi|b Texte imprimé|f Cho
  Hang-bŏm
  210 #|6a01|7ka|a 서울|c 한국문원|d 1997
  210 #|6a01|7ba|a Séoul|c Han'guk munwŏn|d 1997
  700 #|6a01|7ka|a 조|b 항법|f 1958-...|4070
  700 #|6a01|7ba|a Choi|b Hang-bŏm|f 1958-...|4070
Screenshots (10)

doopac : Korean script

AUTHOR 조, 함범 (1958-....)  

TITLE 다시 쓴 우리말 어원이야기 [Texte imprimé] / 조함범  
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海洋渔业经济与渔民社会|f欧阳宗著
200 1#|6a01|7ba|aHaishang renjia|bTexte imprimé|eHaiyang yuye jingji yu yumin shehui|fOuyang Zongshu zhu
210 ##|6a01|7ea|a南昌|c江西高校出版社|d1998
210 ##|6a01|7ba|aNanchang|cJiangxi gao xiao chu ban she|d1998
700 #1|6a01|7ea|a欧阳|b宗著|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因解 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 2#|6a01|7da|a東京|c倉元新社|d1964
210 2#|6a01|7ba|aTōkyō|cSōgen shinsha|d1964
700 3#|6a01|7da|a水野|b清一|f1905-1971|4070
700 3#|6a01|7ba|aMizuno|bSeiichi|f1905-1971|4070
701 4#|6a01|7da|a小|b行雄|f1911-1989|4070
701 4#|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
Tōkyō : Sōgen shinsha, 1964
Multi-script authorities

- **searching Non-Latin forms**

  - 200 #1|aZola|bÉmile|f1840-1902
  - 300 #|aÉcrivain français
  - 400 #|0|5|e|aAgrippa
  - 400 #1|7|ba0yda0y|8|frejpn|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|7|pan|a中世東国の鉄文化解明の前提|bTexte imprimé|f福田豊彦
  200 1#$6a01|7|latn|aChūsei Tōkoku no tetsubunka kaimen 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/raEHf CHj GdEUHY, YHO GdYXje Hf YHO GdhGMO,Sd1198-1256

- **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 : ~ + ^ + o to convert in Unicode precomposed character : dział (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 sh en",
Tōoyō, 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**

  - The library systems don't generally provide input tools.
  - We use IME from Microsoft for the main scripts and we are testing other products, like Keyman.
  - We need to remap the keyboards and specially add some characters to the standard mappings for cataloguing (Cyrillic, Greek...).
  - A visual keyboard is a less efficient solution for productivity.
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.
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.
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