65th IFLA Council and General Conference
August 20-28, 1999

Code Number: 079-155(WS)-E
Division Number: IV
Professional Group: Cataloguing: Workshop
Joint Meeting with: -
Meeting Number: 155
Simultaneous Interpretation: -

Каталогизация средствами универсального набора символов: анализ возможностей

Джоан М. Элипренд
Группа научных библиотек (США)

(Joan M. Aliprand)

Я начала работать как каталогизатор, и хотя сейчас я системный аналитик, я по-прежнему активно интересуюсь данной областью. Когда я изучала каталогизацию в библиотечной школе, первое издание “Англо-американских правил каталогизации”, первых правил, основанных на международных принципах каталогизации, готовилось к изданию. Я думала, что это было последнее слово в каталогизации, немногое можно будет добавить. Как я была неправа! И уж менее всего я думала, что буду вовлечена в этот продолжающийся диалог.

Тема моего сообщения - описательная каталогизация, а главное в нем - пункты, которые принято называть текстом записи. Хотя в фокусе сообщения описательная каталогизация, кое-что может иметь общее применение,т.е. относиться ко всем частям библиографических записей и даже к другим типам библиотечных записей.

В своем выступлении я буду ссылаться на AACR2 [1]. Известно, что сегодня AACR2 применяются не везде. Но поскольку я принадлежу к англо-говорящей среде, я знаю именно эти правила. Кроме того, AACR2 всегда оказывали широкое, как прямое, так и косвенное, влияние . Их прямое влияние проявлялось через перевод на другие языки, чтобы служить в качестве основы для других правил каталогизации. Их косвенное влияние проявляется всякий раз, когда заимствуется любая из множества записей, созданных в англо-говорящем мире. Но даже если английский язык не является языком каталогизации, информация, взятая из источника, может быть полезной и сэкономит время.

Правило 1.0E из AACR2 "*Язык и шрифт описания*" гласит:

В следующих областях описания информация приводится на языке или шрифте (как диктует практика) текста издания:

Область заглавия и сведений об ответственности
Область издания
Область выходных данных
Область серии

Символы и другие знаки, которые невозможно воспроизвести типографскими способами, заменяются соответствующими эквивалентами и приводятся в прямых скобках, что при необходимости оговаривается в примечании.

Главная тема сообщения - транскрипция применительно к новой компьютерной среде, осуществляемая с помощью стандарта Unicode [2] и международного стандарта ISO/IEC 10646 [3].

Эти документы охватывают не только письменности наиболее распространенных языков мира, но и наборы символов и другие элементы текста, такие как математические знаки, шрифт Брайля, пунктуацию и многое другое. Необходимо тщательно следить за тем, чтобы эти наборы символов синхронно обновлялись.

Хочу также остановиться на вопросе правильной транскрипции, т.е.на том, что я называю "точностью" каталогизации.
Далее я упомяну об эффективности поиска, особенно межсистемного поиска, о чем необходимо помнить при принятии каталогизационных решений.

Теперь стало возможным обеспечить автоматизированную поддержку многочисленных шрифтов еще до появления стандартов Unicode и ISO/IEC 10646: в 1983 г. в Информационной сети научных библиотек - RLIN (США) [4] начали применяться стандарты на шрифты, разработанные совместно с OCLC для транскрипции китайского,японского и корейского письма, а стандарты для языков стран Восточной Азии всегда включали несколько шрифтов, но с появлением продуктов, основанных на Unicode, использование многоязычия упростилось.

Стандарты Unicode и ISO/IEC обеспечивают более широкий набор шрифтов и символов, чем это требуется для применения в библиотечной практике, включая USMARC [5] и UNIMARC [6]. Расширение набора шрифтов означает не только такие средства доступа к ним, которых не было прежде, но и включение большего числа символов в существующие шрифты. Ниже приводится сравнительная таблица знаков в нескольких шрифтах.

Шрифт	Категория символа	USMARC/ UNIMARC	JIS X 0208*	Версия 3.0 Unicode
Кирил.	Буквы	102	66	237
Латин.	Дополнит. буквы без знака ударения	21**	0	163
Арабск.	Буквы	124	нет	141
Вост.-Азиатские	Идеограммы	13,469 (86 % из	6,353	27,484

*Японский промышленный стандарт JIS 0208 используется как пример одного из стандартов Восточной Азии. В JIS восточно-азиатские идеограммы называются ***канджи.*** В нем дополнительных букв латинского алфавита нет, но включены латинские буквы A-Z (строчные и прописные), и греческий алфавит.
**Цифра "21" как количество латинских букв без знака ударения в USMARC/UNIMARC является общим числом уникальных специальных знаков в USMARC (American National Standard Extended Latin Character - ANSEL) и UNIMARC(ISO 5426). ANSEL включает 18 специальных знаков; ISO 5426 - 17; 14 знаков являются общими для обоих стандартов.

Не стоит делать выводов, что стандарты Unicode и ISO/IEC 10646 предоставляют все возможности для транскрипции:
(a) Не все, что вы можете увидеть на источнике сведений, есть **в** их наборе.
(b) Не все, что вам необходимо для транскрипции, **может быть** в их наборе.
(c) Некоторые алфавиты требуют наличия дополнительных средств расширенного набора литер для правильного представления.
Но это не означает, что мы должны отказаться от этих стандартов. Я просто хочу, чтобы вы поняли реальность.

Что в них отсутствует

Последние хорошие новости: с добавлением сингальского, эфиопского и монгольского языков, все наиболее распространенные алфавиты мира могут теперь кодироваться. В конце этого года будет издана версия 3.0 стандарта Unicode, а в следующем году намечено к публикации второе издание ISO/IEC 10646.

Списки набора знаков еще не завершены: остаются неохваченными менее распространенные языки, можно добавить еще несколько символов и не решен вопрос с важными мертвыми письменностями, такими как иероглифы и клинопись.(Вряд ли найдется большое количество библиотек, комплектующих и обрабатывающих папирусы и глиняные дощечки, но мертвые шрифты имеют важное значение для науки в целом и для некоторых музеев в частности.)

Единый комплект литер для Unicode даже в его настоящей форме может быть очень большим, и было бы более практичным иметь наборы литер для письменностей, представленных в фондах вашей библиотеки. То, с чем мы чаще сталкиваемся при каталогизации, - это не отсутствие шрифта в целом, а отсутствие определенного символа: например, если в заглавии работы по математике наличествует знак, которого нет в блоке математических операторов. Из-за такой случайности вы не можете обеспечить 100-процентный ввод сведений из источника информации.
Но вы вправе возразить, что вы надеялись найти все необходимые средства в Универсальном наборе символов.
Однако, по разным причинам, вы можете получить ответ "нет".

- То, что вы видите в источнике информации, является чрезвычайно редким символом и его просто нет в таблице;
- То, что вы видите, известно и рассматривается возможность включения в таблицу;
- То, что вы видите, известно, но, согласно принципам построения Unicode, не рассматривается как символ.

Для включения символа в таблицу Unicode особенно важны два принципа: *Символы, но не глифы* и
Унификация через языки. К тому же Унифицированный набор и Классификация иероглифов хэн (“Унифицированный хэн”), разработанная Идеографической согласительной группой*, имеют свои правила обозначения иероглифов.

- Идеографическая согласительная группа (Ideographic Rapporteur Group - IRG) состоит из представителей регионов, где идеограммы используются или имвют культурное значение: Китай, Япония, Корея, Вьетнам, Сингапур, Гон- Конг, Тайвань, а также включает представителя США плюс представителя Консорциума Unicode. IRG отчитывается перед группой ISO/IEC/JTC 1/Sc 2/WG 2, которая ответственна за международный стандарт ISO/IEC 10646.

Принцип *Символы,но не глифы* означает, что некоторые аспекты полиграфического исполнения не имеют значения при определении набора символов. Вот примеры таких аспектов :

- Шрифт *nashki* арабского письма передается стилем *насталик;*
- Способы написания идеограмм в странах Восточной Азии различны;
- В определенных языках способы написания кириллических букв различны;
- Используются сокращения, диграфы и т.д.

Унификация через языки

- Графемы, используемые в определенных языках (или алфавитах), не кодируются по отдельности;
- Различные способы написания букв или идеограмм в различных языках не кодируются по отдельности.

Эти принципы и правила определяют то, что должно кодироваться уникально. В результате , не все, что мы видим в источнике информации, подлежит преобразованию как установленный символ. Это ограничение не является слабым местом стандарта Unicode. Это происходит из-за иного, более сложного видения того, что должно кодироваться в наборе символов.

Первоначальный подход к представлению текста в машиночитаемой форме состоял в том, чтобы передать уникальным кодом каждый отдельный знак на бумаге, хотя уже существовала унификация для наиболее распространенных букв (например, формы строчных латинских букв “a” и “g”). Наборы символов для языков Восточной Азии предоставляли индивидуальные коды для различных способов написания одного и того же иероглифа. В библиотечном наборе символов обычно проявляется тот же подход “кодируй то, что видишь”, за исключением использования знака отсутствия пробела для кодирования латинских букв со знаком ударения, которые кодируются как два символа. (Критики могли бы назвать такую букву “разломленной ”.)

Стандарт Unicode представляет “многослойный” подход к передаче текста. ”Проект для кодирования символов должен обеспечить именно такой набор кодируемых элементов, чтобы дать возможность программистам разработать прикладные программы, способные обеспечить отображение всех текстовых процессов в любом языке”.[9] В результате, помимо другого, символы машиночитаемого текста не соответствуют один к одному визуально читаемому тексту.

Простейший вид представления текста - ***открытый незашифрованный текст*** - простая последовательность кодов и символов. Данные Unicode - открытый текст. Но для более точного отображения необходимо использовать протоколы более высокого уровня, такие как языковую

идентификацию, команды управления форматом, чтобы создать ***усложненный текст*** или ***обогащенный текст.*** USMARC и UNIMARC тоже используют только открытый текст, но их наборы символов предоставляют возможность раздельного кодирования для тех случаев, которые унифицированы стандартом Unicode/ISO 10646.

Итак, необходимо рассмотреть следующие вопросы:

- Насколько точны должны мы быть в процессе транскрипции?
- И если мы должны быть сверхточными, как этого можно достичь при использовании Unicode/10646?

Оценка точности исполнения процесса транскрипции

И вот мы должны рассмотреть вопрос о точности транскрипции. Насколько точной должна быть транскрипция? Почему? Какие исключения мы допускаем (возможно, без сознательного принятия решений)? Какие "побочные средства" мы используем при отсутствии необходимых средств печати?

Точность транскрипции нам необходима для того, чтобы уникально идентифицировать объект описания и сделать его доступным. Однако, заметьте, мы не всегда обеспечиваем транскрипцию на 100 процентов.

Одной из причин неточности транскрипции является то, что правила каталогизации или их интерпретация каталогизирующми учреждениями не всегда требуют и не всегда дают возможность для транскрипции специфических данных. Вот один пример. Иврит на письме не использует гласные,т.е. это письменный невокализированный язык без обозначения гласных и других знаков произношения. Но иногда в источнике информации есть указания на эти обозначения в печатной форме, например, когда автор или издатель хочет, чтобы слово было произнесено необычным образом. Руководство Библиотеки Конгресса по каталогизации на иврите основано на правиле 1.0G, *Ударения и другие диакритические знаки,* и оно интерпретируется (на мой взгляд, неправильно) как запрет транскрипции знаков вокализации, которые присутствуют в источнике информации.

В правиле 1.0E отмечена проблема невозможности достичь точности по причине несовершенства типографских технологий. Но это правило допускает возможность описывать отсутствующие текстовые элементы. Это представляет проблему для межсистемного поиска: следует ли проигнорировать при поиске интерполяцию, или относиться к ней, как к "универсальной вставке", которая соответствует всему, или …? От пользователя едва ли можно ожидать понимания точного описания, составленного каталогизатором.

Существуют также и неписаные правила для подобных отклонений от точности. Описывая антикварные и другие редкие книги, мы автоматически игнорируем особенности литер, каллиграфию и т.д. без всякой попытки зафиксировать эти особенности. Мы исходим из целесообразности, т.к. для большинства современных произведений различия такой степени не имеют значения.

При отсутствии типографских возможностей для передачи всего алфавита существуют различные варианты. Если язык каталогизации использует латинский шрифт, как правило, выбирается процесс романизации: транслитерация или транскрипция оригинального текста знаками латинского алфавита. В 1976 г. Wellish [11] отмечал, что Таблицы Библиотеки Конгресса по романизации (сейчас ALA/LC) были наиболее распространенными, сейчас их заменили таблицы ISO. Если язык каталогизации русский или другой язык кириллической письменности, иногда используется процесс кириллизации. Но не все языки используют

алфавит или слоговую азбуку, и решением может быть перевод информации на местный язык или ведение карточных каталогов по шрифтам.

Все эти альтернативы затрудняют доступ. Если библиотека использует романизацию или кириллизацию, пользователь должен знать об этом, должен знать схему конверсии для каждого языка, чтобы правильно ею пользоваться при определении поискового параметра. Пользователь может не знать о практике библиотеки и использовать совершенно другую схему. Например, при переводах, перевод пользователя может не совпадать с переводом каталогизатора, поиск по карточным каталогам, если они не изданы в форме книг, не может вестись на расстоянии.

Достаточно ли кодированных символов?

Эту проблему можно значительно облегчить через внедрение Unicode/ISO 10646 в USMARC и UNIMARC. Но использование огромного набора кодов не означает, что все можно с точностью транскрибировать. Я сейчас хочу рассмотреть ситуации, когда даже Unicode/ISO 10646 не может обеспечить 100-процентной точности.

Исторически, основной причиной для соблюдения точности при транскрипции было создание эквивалента библиографической сущности с возможно большим количеством деталей. Детали были необходимы, т.к. не было иного пути представить единицу описания на карточке или в книжном каталоге.

Проблемы точности транскрипции обычно связаны с идеограммами, но не только с ними. Если вы каталогизируете звуковую запись, как поступить с символом имени, используемым "артистом, ранее известным как Принц"?

Много сложностей связано с математикой, где двухмерные формулы должны быть втиснуты в одномерное поле. Как представлять математические формулы, используя Unicode, описал Sargent.

Проблемы с идеограммами таковы: либо конкретная идеограмма еще не закодирована, либо варианты формы какой-либо идеограммы представлены одним символом (как отмечали Zhang & Zhen) [12]. Отсутствующие идеограммы включают как действительно уникальные идеограммы (используемые для личных имен), так и те, которые уже используются в специфической среде, но их еще нет в Унифицированном хэне (это иероглифы, официально принятые в Гонконге или используемые в географических названиях). В этой ситуации:

- Отсутствующий иероглиф может быть заменен символом *гета.* Символ *гета* взят из японского книгопечатания и является заменой иероглифа до определения ему символа. Эта техника используется в записях USMARC.
- Символы идеографического описания должны помочь пользователю представить отсутствующий символ. Версия 3.0 стандарта Unicode и второе издание ISO/IEC 10646 включают эти символы.

Если даже специфическая типографская форма будет объединена с другими, а каталогизатор все же хочет использовать именно эту форму, возможны следующие решения:

- Использовать протокол высокого уровня, например, Стандартный универсальный разметочный язык-SGML[13] для представления символа в определенном стиле письма. (Так как USMARC и UNIMARC используют открытый текст, то данная задача за пределами их возможности).
- Представлять идеографические данные в записи, используя шрифт печати, предписываемый кодами языка и названия страны в записи. Например, если код языка

был *chi,* а код страны публикации *cc,* то шрифтом печати должен быть упрощенный китайский стиль. Если код языка - *jpn,* то шрифтом печати должен быть типичный *kanji.* (Эта задача может быть выполнена там, где четкая кодированная информация совместима с предпочтительной формой языка текста и места публикации.)

- Технический комитет Unicode рассматривает предложение, чтобы варианты идеограмм могли указываться в открытом тексте. Может быть, это разрешит вопрос.

Предпочтительные региональные или языковые формы не исключают идеограммы. Если язык урду приводится в арабской письменности, то принято его печатать стилем *nastaliq.* В печати арабский язык обычно передается стилем *nashki.* (Nashki - это стиль печатных литер, используемый в RLIN.) Поскольку вся информация о произведении будет дана одним и тем же типографским шрифтом, то в случае, когда шрифт произведения отличается от шрифта системы, это должно быть оговорено в примечании. Эта ситуация напоминает историю в европейском книгопечатании со староанглийским готическим шрифтом и готическим шрифтом "фрактура".

Общее разрешение проблемы неточной транскрипции в библиографической записи - использование гиперсвязей. В каталоге Web возможно иметь связь с картинкой (сканированным изображением) источника информации. Недостаток сканированного изображения состоит в том, что его невозможно найти из-за специфики конкретной формы глифа, но эта операция более применима к полным текстам, чем к каталогизации.

Заключение

Редакторы правил каталогизации должны пересмотреть правилами транскрипции, чтобы решить, нужны ли изменения в условиях новой технологии. Новая технология позволяет использовать не только стандарт Unicode/ISO 10646, но и вести дистанционный поиск по каталогам через Z39.50.

Тем, кто разрабатывают форматы MARC, необходимо работать вместе с каталогизаторами, чтобы определить, нужна ли переоценка "открытого текста" в действующих форматах. Вопрос заключается не только в том, чтобы принять Unicode/ISO 10646 как утвержденный набор символов (как было сделано с UNIMARC [14] или в деталях рассмотреть необходимые изменения (как делается сейчас для USMARC [15] и для UNIMARC). Это первый и очень важный шаг, хотя требования каталогизации могут идти дальше "обыкновенного текста" стандартов Unicode и ISO/IEC 10646. Если такое требование возникнет, то в различных форматах MARC должна быть предусмотрена специальная методика.

Вопрос, на который необходимо ответить, заключается в следующем: являются ли каталогизационные данные "открытым текстом" или это что-то более затейливое?