CC:AAM Statement in Support of the Internationalization of BIBFRAME
Dec. 13, 2017

The Committee on Cataloging: Asian and African Materials (CC:AAM) strongly encourages an early focus on issues related to internationalization in the development of BIBFRAME, and offers to assist with pursuing that goal. Recognizing that efforts have already begun to address these issues, the Committee submits for consideration the following recommendations to improve global discovery and access to resources described using BIBFRAME.

A checklist for developing internationalization specifications can be found here:

http://www.w3.org/International/techniques/developing-specs?collapse

General considerations

Character encoding:

The full, most recent Unicode character repertoire should always be valid for encoding as UTF-8 in BIBFRAME, subject only to possible "exclusions a priori" such as those currently defined for MARC 21.

https://www.loc.gov/marc/specifications/speccharucs.html#exclusions

Original script and romanization:

To meet the needs of different libraries and cataloging communities and to accommodate both legacy and newly created data, BIBFRAME will need to be able to support a continuing environment of mixed practices including both non-Latin script and/or romanization. Libraries may wish to display the original script of all resources (Latin or non-Latin), or Latin script only (whether original or romanized), or both. In this context it will be particularly important to leave behind our MARC-era assumption that Latin script should always be treated as the default.

Treating all original script (Latin or non-Latin) consistently and coding romanization as a secondary, derived form of transcription will allow flexibility in display and permit targeted provision of access appropriate for different contexts. Original script should be clearly identified, and transliteration into any other script may be coded using subtags (see below). Each instance of transliteration should be clearly linked to its original script (if present) to allow for coordinated display when both are included.

Language tags:

We highly recommend adhering to BCP 47 (Internet Best Current Practice for the use of language tags in cases where it is desirable to indicate the language used in an information object), where possible. Following BCP47 will allow for the greatest possible interoperability with other data on the web. Language tags should be used in lower case.

https://tools.ietf.org/html/bcp47

Consider tagging romanized fields using variant subtags or as per BCP 47 Extension T - Transformed Content.

https://tools.ietf.org/html/rfc6497

Implementation-level considerations:

Character encoding:

A decision should be made on how to handle the byte order mark (BOM) in BIBFRAME: whether to require it, and how to use it.

Original script and romanization:

The treatment of bidirectional text will involve decisions on control characters and markup, and should be considered with reference to UAX #9:

http://www.unicode.org/reports/tr9/

Values for directionality that need to be supported include “ltr”, “rtl”, and “auto”.

The level of rendering support for complex scripts can be expected to vary between browser versions and platforms.

Normalization:

BIBFRAME implementers should consider using Unicode Normalization Form C, following W3C specifications for the World Wide Web and XML.

http://unicode.org/reports/tr15/#Norm_Forms

Language tags:

Most existing MARC data incorporates use of the language codes found in ISO 693-2/B. While the codes in this standard are useful, it may be necessary in implementation to accommodate the codes from ISO 639-1 (2-letter codes) and ISO 639-3 as well. ISO 639-1 covers a subset of widely used languages, while ISO 639-3 allows for much greater specificity in language identification than the other two codes.

https://www.sil.org/x-iso639-3