DWA Assignment

Description of the assignment

The assignment aims to set up a TEI-based framework for presenting editions and metadata of Batak manuscripts, using one manuscript as a text case. To test a linked configuration of the entire set of textual data and manuscript-related metadata, the assignment will involve:

1. Representing the manuscript's metadata in a TEI-compliant structure following the cataloguing practices described in the Beta Masaheft Guidelines <https://betamasaheft.eu/Guidelines/>. It will be attempted to achieve the desired result based on descriptions in my "Catalogue of Catalogues" maintained in a Google Spreadsheet making use of the Google Sheet API <https://developers.google.com/sheets/api>.

2. Diplomatically editing the text of the manuscript edition in TEI-XML, following the EpiDoc Guidelines <https://epidoc.stoa.org/gl/latest/> in general and their specifications for the DHARMA project <https://halshs.archives-ouvertes.fr/halshs-02888186/document> in particular. I will test the degree to which the DHARMA guidelines are applicable to my material, and will attempt to generate not only the diplomatic edition in DHARMA-compliant transliteration <https://hal.archives-ouvertes.fr/halshs-02272407v3> (and to formulate specifications of the DHARMA system that are necessary to do justice to Batak script) but also to generate from the same TEI-encoded data a critical edition in transcription based on standard Batak (i.e., Indonesian) orthography, passing through a script with rules for the conversion from transliteration to transcription.

3. Integrating bibliographic references into the XML file, using the bibliographic data stored and edited in Zotero, and Zotero API <https://www.zotero.org/support/dev/web_api/v3/start>.

4. Using GitHub to maintain the XML data and present it in a static website with GitHub Pages using standard TEI or EpiDoc stylesheets, or those developed for the DHARMA project <https://github.com/erc-dharma/project-documentation/tree/master/stylesheets>.

First Part: Encoding of metadata

The first part of the assignment has focused on the redaction of a XML file, presenting the encoded description of the manuscript Or. 3429 of the University Library at Leiden. For the creation of this first document, I made reference to the guidelines prepared by the BetaMaseheft project, whereas for the structuring of the data in the file, I followed the same order as the one in which the data are presented in my google sheet “Catalogue of Catalogues”. The latter has meant, in particular, presenting first the data related to the repository of which the manuscript is part, followed by its measurements, state of preservation, language and script, topic, line of transmission and, lastly, the relevant bibliographic information and reference.

Second part: the encoding of the *porbuhitan* text

1. Introduction

Before going into the heart of this paragraph, namely describing the framework I have settled upon for TEI encoding of a given Batak manuscript, I would like to make explicit a factor that has helped determine encoding choices I have made. This is the fact that I am co-supervised for my doctoral thesis by Prof. Dr. Arlo Griffiths, who is also one of the principal investigators of the ERC-funded DHARMA project, that focuses on the preparation of encoded editions of texts taken from a wide range of inscriptions and manuscripts from South and Southeast Asia. Thanks to my involvement with Prof. Griffiths, I had the chance to get access to the guidelines prepared for the members of the project, and to make use of its data managing systems on platforms such as GitHub and Zotero.

The corpus at stake in my assignment is not constituted by epigraphic material, but by a single-copy text represented in a tree-bark manuscript, belonging to a genre known in scholarship as *porbuhitan*. In view of the methodological analogies between the study of epigraphic and single-copy text, adhering to the DHARMA guidelines for diplomatic edition (mainly of inscriptions) seemed a natural choice for my project. The EpiDoc-compliant schema prepared for the DHARMA project includes a series of specifications quite adequate to the type of text I aim to encode. Moreover, the DHARMA schema has the advantage of including in the teiHeader of each file important elements, such as the matching patterns for entering bibliographic references into the xml file, which was one of the steps to be completed for my assignment.

2. Sections of the xml file

The first step was to divide the entire file into two big sections, using the element with different values of the attribute @type:

1. → the edition of the text of the manuscript
2. → the bibliographic references

2.1

Where relevant, the element has been further specified with @xml:lang expressing the language used in the original text, as for example xml:lang="bya-Latn", where “bya” corresponds to “Toba Batak”[1].

The internal divisions of the text into different semantic units, are also wrapped in the element but with a different value for the attribute @type, such as @type="textpart". This element is firstly followed by the attribute @n with a numeric value in ascending order, formulated regardless of the different values of type; secondly it is followed by the attribute @subtype, which defines the specific semantic units that will be encoded. The values for this attribute that have been defined for this edition are:

- “introduction”
- “general omen”
- “consequences”

Each of these sections itself consist of one or more elements <p>.

3. Marking up extrinsic structure in the edition

The physical division of the text into pages is marked up with the element <pb> with the attribute @n followed by the number of the page, and the specification r for a recto page, and v for a verso:

- <pb n="1r"/>, etc.

The single lines of text which fill a given page are marked up with the element <lb> modified with the attribute @n followed by the specification of the line number formulated as a siglum, such as “1r/v-x”, where 1r/v- points to the page number, r/v to the side of the page, and the x corresponds to the line number:

- <lb n="1r-1">, etc.

Words split by a line break are marked up with the addition of the attribute @break="no", after the attribute @n.

3.1 Non-alphanumeric signs

Following the rationale that non-alphanumeric elements in the Batak manuscripts do not have solely a decorative function but also have a semantic one, namely to separate the various texts of the manuscript from each other and further to divide a single text into smaller sections, I consider it necessary to mark up the presence of such elements necessary in my edition.

[1] https://iso639-3.sil.org/code/bya

In this regard, the DHARMA schema allows for the element <g> only the attribute @type, with a list of suggested values to define the attribute, describing the features of these elements.

Because of the different functions that non-alphanumeric graphic elements play in Batak manuscripts and since I have already at my disposal a classification of these different elements that can occur in a Batak manuscripts, for each of which I had already defined a specific abbreviation, I suggest the addition in the DHARMA schema of the possibility of having also the attribute @ref for the element <g>.

This different definition of the element <g> requires the specification of the different types of values recognizable for the attribute @ref, which is done within the element <charDecl> in the teiHeader of the file. The use of the element <charDecl> allows the creation of a short authority list of allowed values, defined with the use of the element <glyph xml:id="...">. In this case, the value of @xml:id, will be an acronym that I have defined to refer to the types of non-alphanumeric graphic elements found in the mss. With the creation of this authority list, in the text of the edition the element <g> can then bear the attribute @ref with the values defined in <charDecl>.

3.2 Mark up for interlinear insertions

Any graphemes written, exceptionally, below or above the line of text, will be marked up with the element <add> with the attribute @place. The schema allows the following values for the attribute:

- “inline”
- “below”
- “above”
- “top”
- “bottom”
- “left”
- “right”

3.3 Mark up for variants of graphemes

For encoding the use of an abnormal shape for a given grapheme, two options have been envisaged:

- Use the element <hi> modified with the attribute @rendition, followed by a set of values that have to be defined with an authority list in the teiHeader.
  This approach will yield the following mark up: <hi rendition=”#unified”>na</hi> where “#unified” refers to a given variant of the aksara <na>.

- Use the element <seg> modified with the attribute @type, followed by the value “aksara” and complemented with the addition of the attribute @rendition, followed by a given value, defined, also in this case, in an authority list in the teiHeader.
  This approach will allow the use of the following style of mark up: <seg type="aksara" rendition="#natype2">na</seg>.

I have rejected the first approach since it requires the definition of very specific values, which should describe briefly the features of the variant of the grapheme to be marked up, which would have to be developed first for the Batak script. This approach could be useful for mark up Latin scripts or Western type of scripts for which there is already an standardized nomenclature.

Consequently, I opted for this second approach, which first of all is in compliance with the DHARMA guidelines, and secondly, it has the advantage of being less specific in the mark up. In fact, the attribute @rendition will be followed by values formulated as a siglum, such as “type1” and “type2”. In my encoded file, I have defined under the element <chartDescr>, the values allowed, “type1/type2” and included a short description of the visual appearance of such a variant, enclosed in an element <ab>.

4. Physical Condition and Legibility

4.1 Introductory remarks

Before describing the mark up used for encoding scribal mistakes or illegible parts of the text, I need to underline the importance of a separation, with the use of different mark up, between editorial corrections of scribal mistakes and editorial normalization of unusual spelling used by the scribe.

This has been done following the rationale that, having discerned a recurrent pattern behind unusual spellings used by the scribe, it is plausible that for the scribe himself such spellings were felt to be valid representations of the intended words, so that encoding them as scribal mistakes would have been philologically incorrect.

These unusual spellings are in fact attested in words where a nasal (/n/, /m/, /ng/) is followed by a homorganic consonant.

I will now describe first the mark up used for cases of normalization followed by the mark ups for the editorial corrections.

4.2 Normalization of unusual spellings

These cases have been encoded with the element <choice>, which, according to the TEI schema, groups a number of alternative encodings for the same segment in a text. This element, moreover, requires as children two complementary elements, one option being the element <orig> followed by the element <reg>. The element <orig> marks up the segment as it is found in the original text, whereas the element <reg> marks up the normalized spelling. The use of these paired elements inside <choice> allows the automatic creation of two different editions of the same segment, specifically leaving in the diplomatic edition the spelling used by the scribe, while reserving the normalized reading for another kind of edition.

In order to allow a better understanding I will now furnish two examples of such encoding:

- The word *hompu* (“master”) is recurrently spelt as *hopu* and it has been encoded as <choice><orig>ho</orig><reg>hamo·</reg></choice>pu[2]
- The word *rimbaru* (“new”), when spelt as *ribaru,* has been encoded as: <choice><orig>ri</orig><reg>rami·</reg></choice>baru

For the sake of completeness, I will now list a set of other options that have arisen during my evaluation of this phenomenon, which however I have rejected as inappropriate:

- ho<supplied reason="subaudible">m</supplied>pu

Rejection based on the consideration that such encoding assumes a specific phonetic phenomenon about which I am not sure.

- ho<choice><orig>pu</orig><reg>ma·pu</reg></choice>
- h<choice><orig>opu</orig><reg>ompu</reg></choice>
- <choice><orig>rib</orig><reg>rimb</reg></choice>aru
- ri<choice><orig>b</orig><reg>mb</reg></choice>aru

Rejections based on the consideration that such encodings are not script-oriented, namely do not take into account the structuring of closed syllables with a vowel different from /a/ (see footnote no.2)

[2] In order to allow the reader a better understanding of this encoding choice, it is necessary to briefly discuss an important feature of the Batak script, when dealing with vowels other than /a/ in closed syllables. The writing of these vowels, in fact, is postponed after the last consonant of the closed syllable, before the symbol marking the elision of the inherent vowel /a/ of this latter consonant grapheme. This represent an important factor that I took into consideration for adopting the encoding choice describe in the paragraph since, according to such rule, the expected spelling of /hompu/ is indeed hamo·pu, and consequently it would have been incorrect to mark such cases as a simple omission of the consonant grapheme <m> with the relative sign for the vowel killer <·>.

4.3 Editorial interventions for correction of scribal mistakes

I have distinguished three different categories of scribal mistakes, each of which shall be encoded with specific markup, which I will now list, presenting also the rejected options:

4.3.1 Omission of graphemes

According to the DHARMA guidelines, when one or more characters were erroneously omitted by the scribe, the editor must correct this omission supplying the expected segment, and wrapping it up in the element <supplied>, specified with the attribute @reason, followed by the value “omitted”.

- <supplied reason="omitted">na</supplied>

Moreover, it is important to uphold a clear distinction between cases of editorial correction and editorial restitution, each of which require a specific markup. The rationale suggested by the DHARMA guidelines, is to consider the nature of editorial intervention on the level of phonemes and not on that of characters or glyph components, and to use the two following distinction in markup:

- paro<supplied reason="omitted">·</supplied>taṅan·
- tan<choice><sic>a</sic><corr>o</corr></choice>

Rejected:

- pa<choice><sic>rota</sic><corr>ro·ta</corr></choice>ṅan·
- par<choice><sic>ota</sic><corr>o·ta</corr></choice>ṅan·
- p<choice><sic>aro</sic><corr>or</corr></choice>taṅan·
- ta<choice><sic>na</sic><corr>no</corr></choice>
- tan<choice><sic>a</sic><corr>o</corr></choice>

4.3.2 Superfluous graphemes

Aksara:

- go<surplus>go</surplus>kar

Diacritic:

- panib<choice><sic>i</sic><corr>a</corr></choice>

4.3.3 Substitution of graphemes

hinapatas → *hinapatar*:

- hinapata<choice><sic>s</sic><corr>r</corr></choice>·

4.4 Doubtful readings

Besides these three categories of scribal mistakes, cases of unclearly engraved graphemes have been marked up with the element <unclear>. In the EpiDoc guidelines, the scope of this element is defined as follows: “Character that cannot be identified unambiguously outside of its context”.[3] When the text is only tentatively readable, the element <unclear> can be further defined with the use of the attribute @cert with the value “low”.

4.5 Lacunae

In cases where the written text cannot be read anymore, because of physical damage to the support, the illegible parts have been marked up with the use of the element <gap>. In the DHARMA schema, this element can and must be further defined with the following attributes:

- @reason → with the allowed values “lost”, “illegible”, “undefined”
- @extent
- @quantity → when the exact number of graphemes lost is known, which are expressed with a number as a value of the attribute.
  This attributes, when used, have to be further specified with the use of the following attribute:
- @unit → which has the allowed value of “character”.
  In cases of uncertainty of the precise extent of the lacuna, the attribute @precision can be added with the value “low”.

4.6 Restoring Lacunae

The editorial restoration of physical lacunae, have been wrapped up in the element <supplied>, followed by the attribute @reason specified by the value “lost”:

- <supplied reason=“lost”>na</supplied>

5. Semantic Features

Being aware that these types of tags are optional, both for the DHARMA guidelines that for the TEI in general, and also that my work to deliver the first-ever TEI encoded edition of a

[3] https://epidoc.stoa.org/gl/latest/ref-unclear.html

Batak text represents only a test case, I have limited myself to including three types of semantic markup, namely:

- the element <persName>, with the attribute @type and the values “divine” or “human”
- the element <roleName>, with the attribute @type and the values “datu” or “raja”, further specified depending on the case with the attribute @subtype and the values “guru”, “ompu”, “ompu-tuvan”, “ompu-raja”
- the element <placeName> with the attribute @type and the value “territorialDivision”, further specified with the attribute @subtype and the values “village” or “unknown”

All the values listed for these tags constitute additions to the values allowed in the DHARMA schema.

6. Transformation process: From TEI source to display of the Critical edition in standardized orthography

The TEI encoding of the edition of the *porbuhitan* text is meant to allow generating two documents, a diplomatic and a critical/normalized edition. In order to make this possible, it has appeared necessary to formulate a set of rules, taking into account the features of the original Batak script as rendered in DHARMA-compliant strict transliteration and the corresponding representation in a transcription that relies on the conventions familiar to all Indonesians, namely those of the national “EYD” orthography.

The required transformation should be possible through XLST transformation by relying on the markup in the TEI edition, on the one hand, and on a set of character replacements to be expressed in the stylesheets, on the other. Moreover, it has been necessary to define a precise hierarchy of these automatic transformations, and execute the transformation in two passes.

The first pass should bring about only the tagging of punctuation marks, which in all cases represent editorial interventions into the text, and which thus need to be furnished with appropriate markup. The characters for comma, period and semi-colon are to be replaced by:

- <supplied reason="subaudible">,</supplied>
- <supplied reason="subaudible">.</supplied>
- <supplied reason="subaudible">;</supplied>

A second pass will bring about capitalization or un-capitalization, depending in part on the presence of punctuation signs, and apply also to issues that are unrelated to punctuation.

6.1 Uncapitalization

- All the capital letters I and U, are to be replaced by i and u. In order to avoid conflicts with the automatic capitalization process, and specifically the automatic capitalization of initial letters of <p> and <ab>, or letters following period plus typed space, priority was given to this passage of transformation.

6.2 Transformation of consonants

- The two character ṅ and ṁ should both be replaced by the bigraph ng.
- _k should be replaced by _h (underscore stands for a typed space)
- hV (V stands for any lowercase vowel) should be replaced by V (the V should be lowercase except if the word stands at beginning of <p> or of a new sentence — see capitalization rules below)

6.3 Capitalization

- capitalize every initial letter of a name enclosed in the element <name>, nested under the element <persName>
- capitalize every initial letter of a <p> or <ab> and every letter following ._ (period plus typed space)

6.4 A more complicated replacement pattern

The most tricky element for this process of transformation of the xml file is related to a specific feature of the Batak script which has required the creation of an operational string, in order to have an automatic redaction of the critical/normalized edition.

The strategy I applied has been representing in the xml file the way in which the graphemes have been written in the text, and entrusted the transformation needed for the editorial/normalized edition to an automatic processing of the file. In order to have this automatic transformation, I design the following regular expression to be added to the stylesheet:

- (a)(\w)([eiou])(·)

Where (a) points the inherent vowel /a/ of the first grapheme, (\w) refers to a variable consonant grapheme, the ([eiou]) points to the different vowels that can be found after the second graphemes and lastly the (·) is the symbol used for representing the vowel killer sign used in the Batak script. This regular expression will help the machine find in the encoded text this combination of vowel, consonant and median dot and, successively, the definition of a replacement pattern, will substitute the first vowel /a/ with one of the possible vowels mentioned in the square brackets, furnishing finally in the critical edition the intended word with its appropriate spelling. The replacement pattern has been formulated as follows: $3$2.

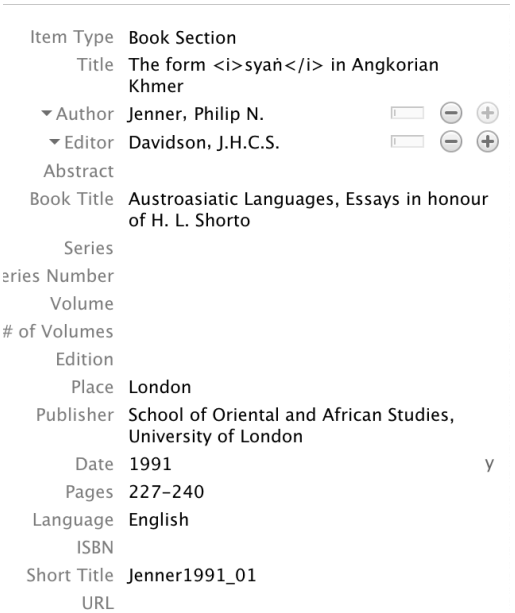
Third Part: integration of the manuscript metadata and bibliographic reference

1. Introduction

The third part of the assignment has focused on the integration of bibliographic data, related to the manuscript or to the text, into the XML file. Also in this case, I have applied the workflow suggested by the DHARMA guidelines, which is based on an integration between the bibliographic data stored in a common group library managed through Zotero, and the XML file itself. The workflow suggested is split into two parallel phases, one related to a regulated structuring of the bibliographic dataset on the group library on Zotero, and a second related to the encoding of these data inside the XML file of the edition of the text.

2. Workflow in Zotero

As for the first phase, each of the bibliographic entries in the Zotero library has to be furnished with a unique internal identifier, to be stated in the field “Short Title” of the Zotero entry. This identifier is created as an abbreviation including the Surname of author, year of publication and, in order to be able to differentiate multiple references for the same author in the same year, an extension with a number in ascending order, separated from the previous two elements by an underscore. Example: Jenner1991_01/02 ...



Item Type	Book Section
Title	The form syaṅ in Angkorian Khmer
Author	Jenner, Philip N.
Editor	Davidson, J.H.C.S.
Abstract	
Book Title	Austroasiatic Languages, Essays in honour of H. L. Shorto
Series	
eries Number	
Volume	
# of Volumes	
Edition	
Place	London
Publisher	School of Oriental and African Studies, University of London
Date	1991
Pages	227–240
Language	English
ISBN	
Short Title	Jenner1991_01
URL	

3. Workflow in the XML file

As for the second phase, in order to allow automatic connection between the Zotero library group and the XML file, in the DHARMA template, the teiHeader of the file already includes the element <listPrefixDef>, which, according to the TEI guidelines, “contains a list of definitions of prefixing schemes used in teidata.pointer values, showing how abbreviated URIs using each scheme may be expanded into full URIs,”[4] followed by the element <prefixDef>, with the addition of the attribute @matchPatter="([a-zA-Z0-9\-_]+)" and @replacementPattern="https://www.zotero.org/groups/1633743/erc-dharma/items/tag/$1">.
The workflow then requires the addition of the encoded bibliographic references in the section of the file , which will contain the following elements in succession:

1. <listBibl> with the attribute @type, distinguished with the two values “primary” and “secondary”.
2. <bibl> with the attribute @n that can be used to define a siglum for a previous editor. Example: <bibl n=”PJ”>. This siglum PJ can then be displayed if any variant reading by that scholar are to be reported.
3. <ptr> with the attribute @target. This last attribute will be specified, case by case, with a specific value, structured as “bibl:Jenner1991_01”, where the segment after “bibl: … ” points to the abbreviation of the bibliographic entry as created in the field “Short Title” in Zotero. Example: <ptr target=”bibl:Jenner1991_01”/>.

In addition, the inclusion of the specific locus within the bibliographic entry, such as the page (range), or identification number of a manuscript or of an inscription, has to be encoded with the attribute <citedRange> specified by the attribute @unit, which has the following allowed values:

- “page” for page numbers
- “part” to distinguish sections of publications where identical page numbers re-occurs within a single volume
- “volume” for multi-volume publications
- “note” for a numbered foot or end note
- “item” for a number in an anthology of editions,or an item in a numbered list
- “entry” for an entry in a dictionary or encyclopedia
- “figure”, “plate”, “table”, “appendix” for visual material

Examples: <citedRange unit="page">122-156</citedRange>;
<citedRange unit="item">XIII</citedRange>

[4] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-listPrefixDef.html

4. Transformation to HTML file

In conclusion, thanks to the specification of the matching pattern of the Zotero URLs files in the teiHeading of the XML file, and with the encoding of the bibliographic reference in the XML file, nested as described above, and especially with the element <ptr> and the attribute @target followed by the siglum of the bibliography entry as stored in Zotero, the process of transformation of the file will automatically recall the bibliographic reference stored on Zotero and present it in its extended version in the final version of the file in HTML format.