Using TEI XML Schema to Encode the Structures of Sarawak Gazette

Tze-Min Fong and Bali Ranaivo-Malançon

***Abstract*—Automatic extraction of information from old printed documents which have been digitised injudiciously will end up with a lot human corrections. To overcome the problem, one possible solution is to annotate the documents with some markups. This paper presents the encoding of the digitised sample of Sarawak Gazette published from 1903 until 1939 using the standard TEI XML schema. The output of the work is a set of six TEI XML templates that is considered to represent the different layout structures found in the studied samples.**

***Index Terms*—Data structure, layout analysis, metadata, TEI P5 schema.**

I. Introduction

Sarawak Gazette is one of the oldest newspapers published in Sarawak. The first publication was on Friday, August 26, 1870. This old newspaper contains a lot of interesting information, and has become an essential source of historical information of Sarawak events, such as trade and economic activities, law and order, agriculture information, mineral and oil production statistics, anthropology and archaeology, etc. Extracting information depicted in Sarawak Gazette will help certainly the preservation of the history of Sarawak. However, a direct extraction is limited due to the fact that in general, the information is in unstructured form. Thus, adding some markups that identify clearly and without ambiguity the different components of Sarawak Gazette will facilitate the retrieval of information.

To encode the information, the layout of Sarawak Gazette needs to be studied and determined formally, and then a metadata structure based on the layout studies can be designed properly. In this work, the metadata structure is based on the Text Encoding Initiative (TEI) latest guidelines, TEI P5. The overall process is illustrated in Fig. 1 and these steps will be followed as the structure of this paper.

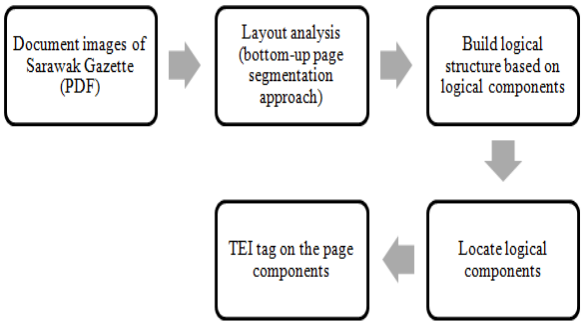

Fig. 1. Sarawak gazette metadata design process flow.

Manuscript received June 12, 2014; revised August 14, 2014.

Bali Ranaivo-Malançon is with the Universiti Malaysia Sarawak, Malaysia (email: mbranaivo@fit.unimas.my).

The process starts with the original document images of Sarawak Gazette. The document images should be in PDF format. Then, the PDF documents images will be converted to JPEG image, and undergo layout analysis by using the bottom-up page segmentation approach. Once the layout structure of Sarawak Gazette is detected, a logic role can be associated to some of its components. The logical components will be arranged in a hierarchical structure, which is called logical structure. It describes the relationship between logical components, for example, a document includes title, authors, summary, and a sequence of chapters. A chapter might include a title, and a sequence of sections, and so on.

Subsequently, the logical components can be located and tagged by TEI by matching the layout structure of each page of document images against models of logical components.

II. Importance of Metadata Structure on Sarawak Gazette

Other than facilitating the information extraction from Sarawak Gazette, metadata structure plays a crucial role in the Sarawak Gazette digitization, OCR and linguistic processing in the possible future. Sarawak Gazette has large amount of scanned pages and very bulky, and metadata is essential to manage and control over the large amount of items. Metadata will guide the process of digitization, in terms of evaluation and quality control. It also helps to make sure that the digitized data are accessible, sustainable and integratable.

III. Sarawak Gazette as Scanned Documents

Sarawak Gazette is one of the oldest newspapers published in Sarawak with the first publication on August 26, 1870 by the Government Printing Office. It was initiated by Charles Brooke, the first White Rajah of Sarawak. The objectives were to provide Europeans who live at outstations, concise statements of official business and other issues of public interest, and to serve as an official report of the condition of the various residencies under the Sarawak Government. It was published monthly to play the role as newspapers which edited by the Rajah's Civil Service [1], [2].

The publications of Sarawak Gazette from 1870 until January 1, 1984 have been scanned and stored in PDF image files. However, the proposed metadata structure of this project will cover only the contents of Sarawak Gazette from publication year 1903 until 1939. The scanned documents are not in a very good condition (Fig. 2).