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.

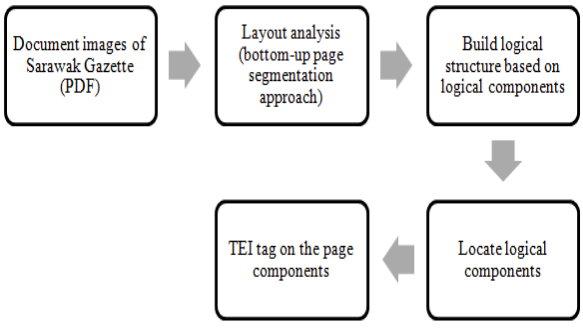


Fig. 1. Sarawak gazette metadata design process flow.

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).

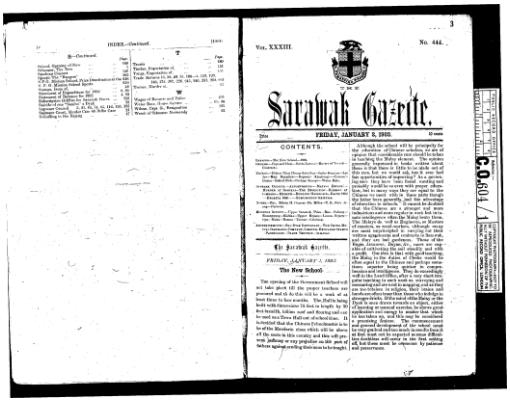
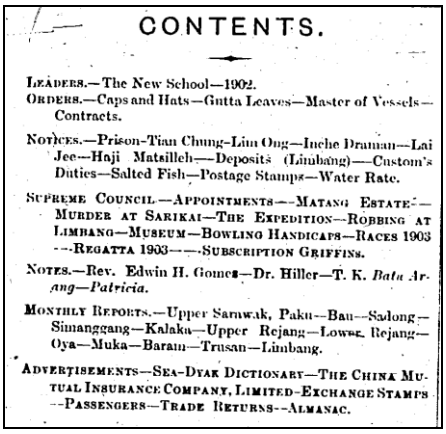


Fig. 2. A sample of the scanned Sarawak gazette.

Even a very powerful optical character recognition (OCR) like ABBYY FineReader fails to produce a readable text as shown in Fig. 3 (the original scanned newspaper) and Fig. 4 (the output of the OCR).



CONTENTS.

LEADERS.—The New School—1902.
ORDERS.—Caps and Hats—Gutta Leaves—Master of Vessels—Contracts.
NOTICES.—Prison-Tian Chung-Lim Ong—Inche Draman—Lai Jee—Haji Matsilleh—Deposits (Limbang)—Custom's Duties—Salted Fish—Postage Stamps—Water Rate.
SUPREME COUNCIL—APPOINTMENTS—MATANG ESTATE—MURDER AT SARIKAI—THE EXPEDITION—ROBBING AT LIMBANG—MUSEUM—BOWLING HANDICAPS—RACES 1903—REGATTA 1903—SUBSCRIPTION GRIFFINS.
NOTES.—Rev. Edwin H. Gomes—Dr. Hiller—T. K. *Batu Ar-ang—Patricia.*
MONTHLY REPORTS.—Upper Sarawak, Paku—Bau—Salong—Simanggang—Kalaka—Upper Rejang—Lower Rejang—Oya—Muka—Baram—Trusan—Limbang.
ADVERTISEMENTS—SEA-DYAK DICTIONARY—THE CHINA MUTUAL INSURANCE COMPANY, LIMITED-EXCHANGE STAMPS—PASSENGERS—TRADE RETURNS—ALMANAC.

Fig. 3. A portion of the scanned Sarawak Gazette in Fig. 2.

CONTENTS.
Leaders. - The New School - 1902.
Orders. - Caps and Hats - Gutta Leaves - Master of Vessels - Contracts.
Notices. - Prison-Tian Chung-Lim Ong - Inche Draman - Lai Jee - Haji — Deposit* (Lim\>a'n<;}—Custom's
Duties—Salted Fish—Postage Stump*—Water Kate.
Hipkeme Council—Appointments—Matan<; Estate—murder at sarikai—tlle expedition'—ropbixg at Limbano—Museum—Bowlisq
Handicaps—Races 1903 —Regatta 1903——Subscription Griffins.
Notes.—Rev. Edwin Il. flomei—Dr. Hiller—T. K. Itatn .lr-ting—Patricia.
Moxtiily Rf.kh-.ts.—Jppci Sarawak, Paku—B&u—Sailonj; —
Simanssang—Kalaku—Upper Rejanx —l.o»«. liejan^'— O.va—Muka—
Baram—Trusan—Limbang.
Advertisements—Sea-Dtak Dictioxart—The China Mutual Insurance
Company. Limited-Exchange Stamps —Passengebs—Trade Returns—
Almanac.

Fig. 4. OCR output of Fig. 3.

Thus, to assist an automatic recognition of the characters in Sarawak Gazette, it is important that the layout structure of the newspaper is identified first.

IV. Document Analysis of Sarawak Gazette

In order to markup the different component parts of each page of the Sarawak Gazette, it is essential that a document analysis has to be carried out. There are many techniques in extracting and analysing the layout of a document. For example, [3] has proposed a bottom-up page segmentation approach for the layout analysis of an Arabic newspaper. The developed algorithm takes into account the complex structural layout of Arabic newspaper. The algorithm is based on the connected components for image, thread, and frame extraction. For the case of Sarawak Gazette, the document analysis is done in two successive steps that are layout identification and then layout analysis. The layout extraction was done automatically whereas the layout analysis was carried manually.

A. Automatic Layout Extraction

Two image processing software were chosen for the layout extraction of some selected samples of Sarawak Gazette: SCRIBO Document Layout Analysis and Reconstruction and Fiji Image Processing Package.

SCRIBO (Semi-automatic and Collaborative Retrieval of Information Based on Ontologies)[1] is a research project that aims to provide algorithms and free software for annotating semi-automatically and collaboratively digitised documents. The proposed method is based on the automatic knowledge extraction found in texts or images. SCRIBO has a tool, which is available online and dedicated to the layout analysis of historical documents. The selected samples of Sarawak Gazette have been submitted to SCRIBO. The automatic layout analysis by SCRIBO of the example of Sarawak Gazette in Fig. 2 is shown in Fig. 5.

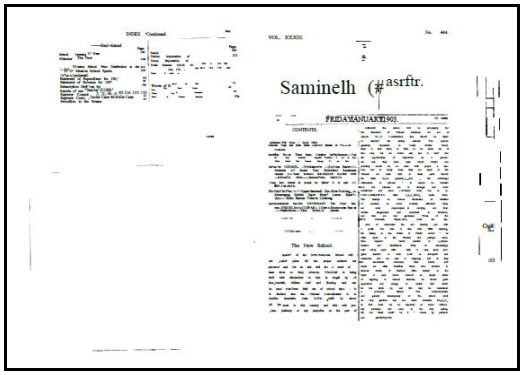


Fig. 5. Layout extracted by SCRIBO of Fig. 2.

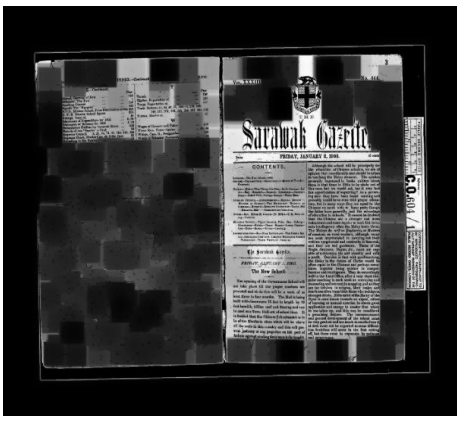


Fig. 6. Layout extracted by Fiji of Fig. 2.

Fiji [2] is an image processing package that has been developed to assist research in life sciences. It offers a large

[1] SCRIBO Historical Document Layout Analyis, https://olena.lrde.epita.fr/demos/historical_document_layout_analysis.php.
[2] Fiji is Just ImageJ, http://fiji.sc/Fiji.

variety of plugins and image processing features. It is simple, easy to use, and can be installed directly on personal computer. The Fiji feature called "Subtract Background" has been used to process the selected sample of Sarawak Gazette. Fig. 6 shows the result of this process on the file in Fig. 2.

SCRIBO and the Fiji tool used for layout extraction require the input file to in one of image formats. Thus, the original Sarawak Gazette PDF files were converted into JPEG. The different file formats involved in the conversion process is presented in Fig. 7.

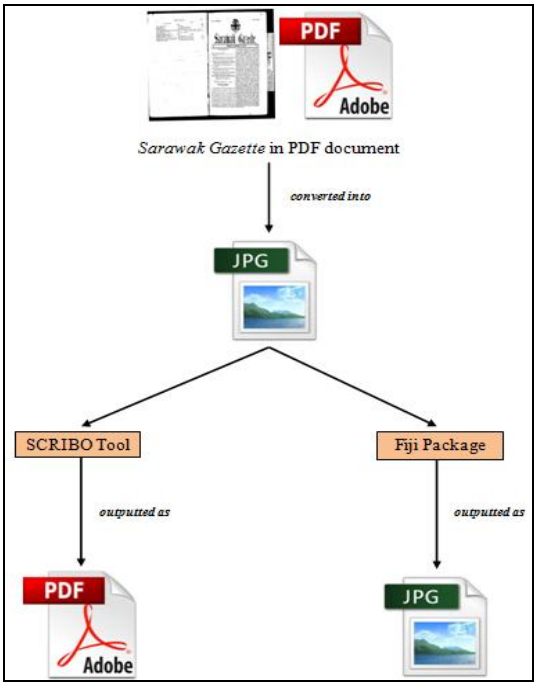


Fig. 7. File format conversion during layout extraction.

B. Manual Layout Analysis

By mapping the SCRIBO and Fiji outputs, a complete layout and structure of two-pages of Sarawak Gazette can be visualized clearly. A layout analysis is a sketched diagram that represents all the components of a given document. Fig. 8 shows the diagram of the layout analysis of Fig. 2 based on the human interpretation of Fig. 5 and Fig. 6.

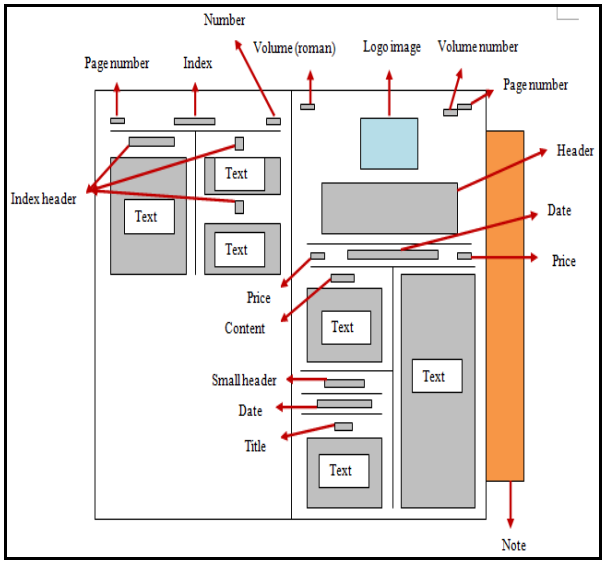


Fig. 8. Layout analysis from Fig. 5 and Fig. 6.

V. TEI as a Standard Metadata for Sarawak Gazette

As defined by NISO in their document "Understanding Metadata" [4], "metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource." Another straight forward and very popular definition of metadata is "data about data". A metadata describes the attribute of a document or object. The concept of metadata is very important for librarians, authors, digital archivists, database developers, or end users who are searching information in electronic documents. Reading a document which has been increased with metadata structure is not meant for humans but rather for machine. However the document is still intelligent for human.

The three main types of metadata are descriptive metadata, structural metadata, and administrative metadata. A descriptive metadata "describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords" [4]. A structural metadata "indicates how compound objects are put together, for example, how pages are ordered to form chapters" [4]. An administrative metadata "provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it" [4]. The encoding of Sarawak Gazette is more a structural metadata. According to [5], a good metadata structure extends the use of data by researches and to prolong the data life span.

A. TEI P5 Guidelines

TEI stands for 'Text Encoding Initiative'. It is a consortium dedicated to the development and maintenance of a standard for the representation of texts in digital form. The latest guideline is TEI P5 [6]. The encoding schema is formulated as an application of the Extensible Markup Language (XML) using UNICODE as an international standard of character encoding. A TEI document based of the P5 guideline must be expressed as a valid XML-conformant document which uses the TEI *namespace* appropriately.

TEI P5 Guidelines contain recommendations on the appropriate ways to represent the features of textual resources which need to be identified explicitly so that they can be processed by computer programs. These guidelines state a set of tags which may be inserted in the electronic representation of the text to represent the text structure and other interesting features. The guidelines are applicable to text in any natural language, of any date in any literary genre or text type without any restriction on form or content [6].

As stated by [7], TEI is a well-understood format for lasting preservation of digital information and metadata. TEI Guidelines is the chief deliverable which specified encoding methods for machine-readable texts. It marks up the electronic text such as novels, plays, and poetry [4]. It focus on the exchange of textual information, and applicable in the creation of new resources and in the interchange of existing ones.

However, according to [8], TEI is very complex, if it is compared to other similar element tagging, such as Dublin Core, METS and MARC, but the complexity is inherent in the task of annotating historical document. Reference [8]

illustrates the encoding of a primary source using TEI. The source's name is *Alteres Landbuch*. It is a code of law and customs from the Swiss canton of Appenzell which originally created around 1450. However, it has been modified in later years. Encoding such a text in TEI is not a machine-driven process, but it depends in many cases on the paleographical, linguistic, and historical interpretation of the text, since the manuscript is relatively old [8].

B. TEI Modules, Elements and Attributes for Sarawak Gazette

The suitable TEI elements for the metadata structure of Sarawak Gazette are determined based on the components and logical structure that have been identified during the document analysis presented in Section III.

The TEI P5 encoding schema consists of 21 modules (analysis, certainty, core, corpus, dictionaries, drama, figures, gaiji, header, iso-fs, linking, msdescription, namesdates, nets, spoken, tagdocs, tei, textcrit, textstructure, transcr, and verse). Each module declares particular XML elements and their attributes. Table I shows the selected modules and elements used to design the metadata structure of Sarawak Gazette. The meaning of each element can be found in the TEI P5 Guidelines [6].

TABLE I: TEI MODULES AND ELEMENTS FOR SARAWAK GAZETTE

Module Name	Elements
core	addrLine; cb/; date; graphic; head; item; l; lb/; list; measure; note; p; title
figures	cell; row; table
header	fileDesc; publicationStmt; sourceDesc; teiHeader; titleStmt
textstructure	back; body; div; front; signed; text
transcr	fw

Not all elements in Table I have attributes. Table II lists the TEI elements that make use of attributes in the Sarawak Gazette metadata design.

TABLE II: TEI ELEMENTS AND ATTRIBUTES FOR SARAWAK GAZETTE

Elements	Attribute Classs	Attribute
cb	att.global	@n
div	att.typed	@type
fw	-, att.placement	@type, @place
graphic	att.resourced	@url
head	att.global	@rend
measure	-	@type
note	att.placement	@place
table	-	@rows, @cols

VI. Creating Layout Template in TEI for Sarawak Gazette

Once the TEI elements and attributes have been defined, the next step is to create the layout template for each identified unique layout structure of the Sarawak Gazette. Based on the document analysis step, it has been found that the whole set of Sarawak Gazette, from 1903 until 1939 can be represented by six templates only. It means that the layout structure of each published page of the Sarawak Gazette during the given period is one of the six templates.

TEI, as a very active community, has pre-loaded TEI templates in <oXygen/> XML Editor [9], which is a commercial software. This capability of <oXygen/> has simplified a lot the manual task of creating the TEI templates for Sarawak Gazette. Thus, the XML files for Sarawak Gazette were created using <oXygen/> XML Editor with the TEI DTD and stylesheets which support TEI P5 version.

Because Sarawak Gazette has been digitised with two printed pages in one scanned page, each created TEI template represents the two printed pages. Fig. 9 shows one of the data structure design output in TEI XML document.

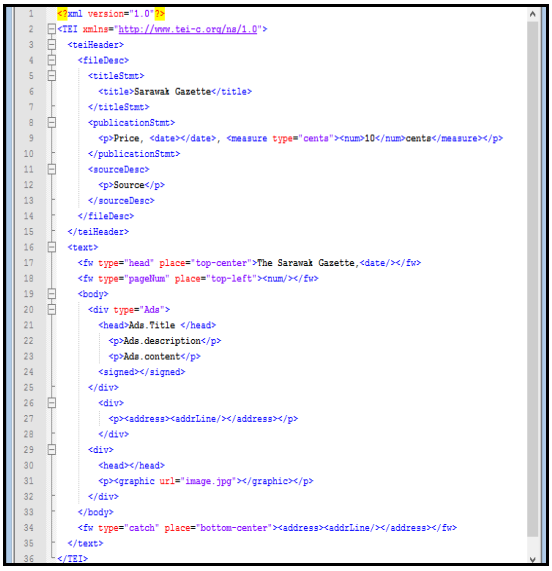


Fig. 9. Layout template in TEI of a sample of Sarawak gazette.

VII. Result Analysis and Validation

TEI Guidelines authorize the use of well-formed XML as representation format. XML documents must obey to the World Wide Web Consortium recommendation of the XML 1.0 (Fourth Edition). The design of metadata and data structure of Sarawak Gazette based on TEI P5 Guidelines is expressed in XML document. The XML documents of the metadata design used the TEI namespace with proper declaration at the beginning of the TEI text: <TEI xmlns= http://www.tei-c.org/ns/1.0/>.

All the XML documents of the metadata design match all the rules and syntactic constraints defined by the TEI P5 Guidelines. No syntax error has been noticed in all the XML documents. The markup in the document is accurately represents the TEI abstract model. Other than that, all the documents of the metadata design contain a single <teiHeader> element which includes elements for Title Statement, Publication Statement, and Source Statement. It is a mandatory component of a valid TEI XML document. Metadata and data structure design in this research are suitable for most of the possible layout in Sarawak Gazette from year 1903 to 1939. The samples chosen are from publication of the year 1903, 1913, 1919, and 1921 which covered the different layout changing along those years.

VIII. Conclusion and Future Works

The result of the layout analysis and the metadata design of Sarawak Gazette in TEI XML documents is the first step

before exploring automatically the contents of the newspaper. All the components identified during the Sarawak Gazette document analysis can be tagged by TEI elements with correct meanings as defined in the TEI P5 Guidelines. The structural relationship of the layout is described correctly by the TEI text structure elements.

The TEI XML templates of Sarawak Gazette are still under reviewed and opened for discussion. They will be adopted fully if they get the approval of a variety of users such as librarians, historians, document analysts, etc.

As highlighted in [4], "metadata is [a] key to ensuring that resources will survive and continue to be accessible in the future".

ACKNOWLEDGMENT

This research is supported by UNIMAS Short Grant Scheme through Grant no. 02(S103)/875/2012(16) to Bali Ranaivo-Malançon. The authors would also like to thank the Universiti Malaysia Sarawak for providing the resources used in the conduct of this study.

REFERENCES

[1] P. N. Sarawak. (2013). *E-Sarawak Gazette*. White Hornbill. [Online]. Available: http://www/pustaka-sarawak.com/gazette/about_us.php.
[2] *Sarawak Gazette Delima Edition*. (2006). Faradalemedia.com. [Online]. Available: http://www.faradalemedia.com/sg/home.html.
[3] K. Hadjar and R. Ingold, "Arabic newspaper page segmentation," in *Proc. the 7th International Conference on Document Analysis and Recognition (ICDAR' 03)*, USA, 2003, vol. 2, pp. 895.
[4] National Information Standards Organization (NISO), *Understanding Metadata*, USA, Niso Press, 2004.
[5] K. Beard, "A structure for organising metadata collection," in *Proc. the 3rd International Conference/Workshop on Integrating GIS and Environment Modeling*, Santa Fe, 1996.
[6] L. Burnard and S. Bauman, "TEI P5: guidelines for electronic text encoding and interchange," Oxford, TEI Consortium, 2011.
[7] J. Cummings. (Sept. 2013). *An Inroduction to the Text Encoding Initiative (TEI)*. [Online]. 8. Available: http://prezi.com/s8rqk-xdpzdb/an-introduction-to-the-text-encoding-initiative/
[8] M. Piotrowski, "Natural language processing for historical texts," *Synthesis Lectures on Human Language Technologies*, vol. 5, 2012.
[9] Visual TEI Editor. (2002). <oXygen/> XML Editor. [Online]. Available: http://www.oxygenxml.com/xml_editor/tei_editor.html

Tze-Min Fong was born in Kedah, Malaysia on March 6, 1990. She is currently persuing her bachelor degree of computer science with Honors at Universiti Malaysia Sarawak in 2014. She worked as an annotator for a project on *Sarawak Gazette*. She is currently doing her research on designing metadata structure for *Sarawak Gazette* based on Text Encoding Initiative Guidelines, as her final year project at Universiti Malaysia Sarawak.

Bali Ranaivo-Malançon was born in madagascar. She gained her PhD in NLP from the national institute for oriental languages and civilizations (INALCO, Paris, France) in 2001. She is currently an associate professor at the Universiti Malaysia Sarawak (UNIMAS, Malaysia). Her research interests are geared towards the development of linguistic resources, text proccesing, text mining, and processing of historical documents.