Pendokumenan dan Pembangunan Korpus bagi Bahasa Peribumi di Sarawak melalui Kaedah Pendokumenan Komputasi Bahasa

oleh
Suhaila Saee
Pensyarah Kanan
Fakulti Sains Komputer dan Teknologi Maklumat
Universiti Malaysia Sarawak

Pengenalan

Populasi penutur bahasa asli di seluruh dunia semakin berkurangan. Penurunan jumlah ini disebabkan oleh tekanan bahasa global yang dominan (seperti bahasa Inggeris yang merupakan bahasa lingua franca perdagangan antarabangsa, penyelidikan, dan Internet), migrasi luar bandar-bandar, dan eksogami (kumpulan antara etnik perkahwinan). Begitu juga, jumlah penutur 63 bahasa Sarawak juga menurun. Oleh itu, Kumpulan Penyelidikan Teknologi Bahasa Peribumi Sarawak (SaLT) di Universiti Malaysia Sarawak telah memulakan bilangan projek penyelidikan dan pembangunan dengan tujuan akhir untuk menghidupkan semula dan mengekalkan bahasa pribumi Sarawak. Projek yang sedang dijalankan merangkumi pembinaan korpus bahasa (Iban, Melanau, dan Kelabit), serta, penyelidikan dan pengembangan teknologi yang menyumbang kepada pelaksanaan perisian untuk bahasa etnik. Secara khusus, projek-projek ini merangkumi pengembangan morfologi penganalisis dan penanda Part of Speech (POS) yang mempunyai pengetahuan mengenai terjemahan bahasa Iban-Inggeris, dan dalam interaksi komputer manusia menggunakan ucapan dan teks Melanau. Projek lain dalam 'pipeline termasuk wiki pendekatan dalam membina leksikon Bidayuh, dan kamus bahasa Melayu Sarawak berasaskan web. Melalui inisiatif dan projek, pendekatan pelbagai disiplin untuk memelihara bahasa peribumi sedang dikembangkan dan diperhalusi.

Keperluan Pendokumentasian Sumber Bahasa

Kaedah konvensional untuk mendokumentasikan kepelbagaian sumber bahasa melibatkan pertuturan dan teks adalah sangat bergantung pada kepakaran dan usaha ahli bahasa di peringkat linguistik. Proses sedia ada ini memakan masa, usaha keras ahli bahasa dan melibatkan kos yang tinggi. Walau bagaimanapun, kebanyakan kaedah sedia ada dikembangkan hanya untuk bahasa sumber tinggi seperti Bahasa Inggeris dan bergantung pada sejumlah besar data ucapan dan teks. Dengan status semasa, kebanyakan bahasa dunia tidak mempunyai sumber bahasa seperti itu, maka penerapan teknologi bahasa untuk dokumentasi bahasa komputasi adalah terhad.

Oleh itu, perbincangan ini bertujuan untuk membangunkan korpora sebagai strategi untuk mendokumentasikan, menganalisis, mengarkibkan dan merevitalisasi bahasa yang kurang sumber di Borneo dengan sistem cadangan *(recommendation system)* menggunakan pendekatan sumber orang ramai *(crowdsourcing)* dan kolaboratif. *Crowdsourcing* didefinisikan sebagai istilah payung untuk pelbagai pendekatan yang memanfaatkan potensi orang banyak dengan menyebarkan panggilan terbuka untuk sumbangan dalam sesuatu tugas (Geiger et al, 2012). Menurut *Oxford Handbook of Endangered Languages* oleh Regh dan Campbell (2018), pendekatan kolaboratif untuk dokumentasi bahasa adalah 'untuk menangani pelbagai keperluan semua pihak yang berkepentingan dengan cara yang bertanggungjawab, timbal balik dan hormat (Rice, 2006) akhirnya mengaburkan garis antara 'penyelidik' dan 'subjek'. Kedua-dua pendekatan tersebut melibatkan komuniti maya, baik penutur asli atau masyarakat, untuk turut serta dalam membangun korporat bahasa Borneo yang kekurangan sumber bahasa. Karya ini diharapkan dapat mendokumentasikan bahasa-bahasa di bawah sumber Borneo yang membantu melestarikan dan menghidupkan kembali bahasa-bahasa tersebut dalam bentuk digital dan sebagai platform pembelajaran untuk masyarakat dan generasi muda untuk mempelajari lebih lanjut mengenai bahasa-bahasa tersebut.

elektronik-Cerita Rakyat Borneo, eCERAB

Dewan Bahasa dan Pustaka Cawangan Negeri Sarawak (DBP-Sarawak) mempunyai koleksi cerita rakyat pelbagai suku kaum yang masih dalam bentuk kaset sebanyak 951 judul keseluruhan yang ada dalam simpanan untuk proses pendigitalan. Pihak UNIMAS, terutamanya, Fakulti Sains Komputer dan Teknologi Maklumat telah bekerjasama dengan pihak DBP-Sarawak di dalam usaha mendigitalkan kaset-kaset ini pada tahun 2017. Projek ini telah berjaya membina data digital cerita rakyat Borneo daripada 357 buah kaset sebanyak 816 judul cerita. Data digital ini memerlukan sistem pengurusan yang efisien dimana ia boleh

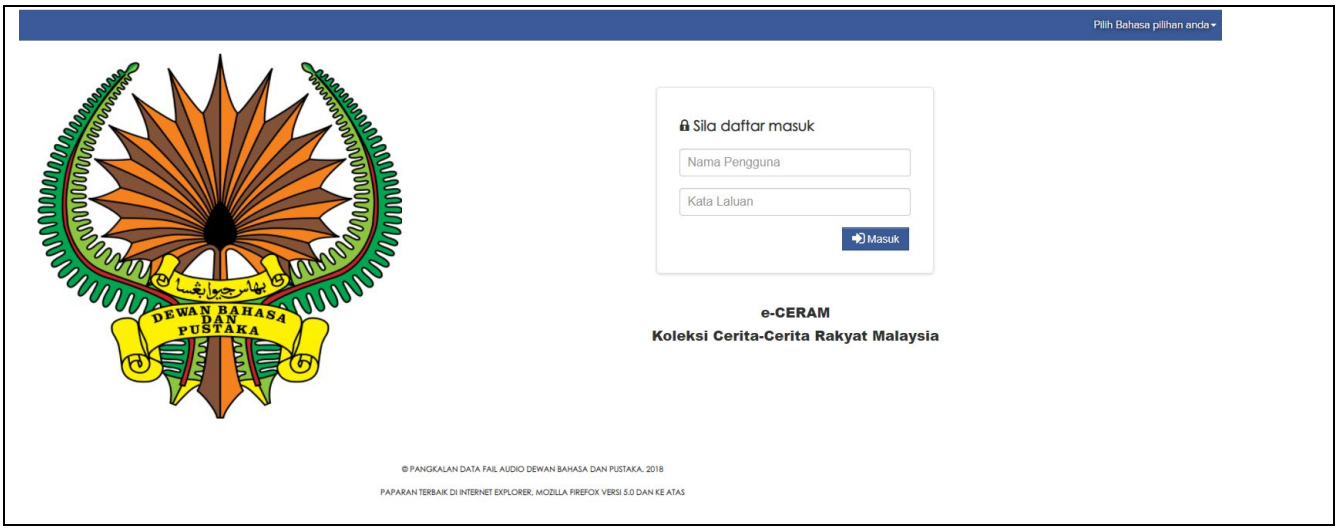
membantu staf DBP mengurus dan membina data transkripsi. Maka dengan itu, DBP bekerjasama dengan UNIMAS di dalam menghasilkan sistem pengurusan data cerita rakyat Borneo atau e-CERAB.

e-CERAB merupakan satu sistem pengurusan data berasaskan web yang dibangunkan untuk menguruskan data audio yang telah disalin semula dari kaset kepada format audio, MP3.

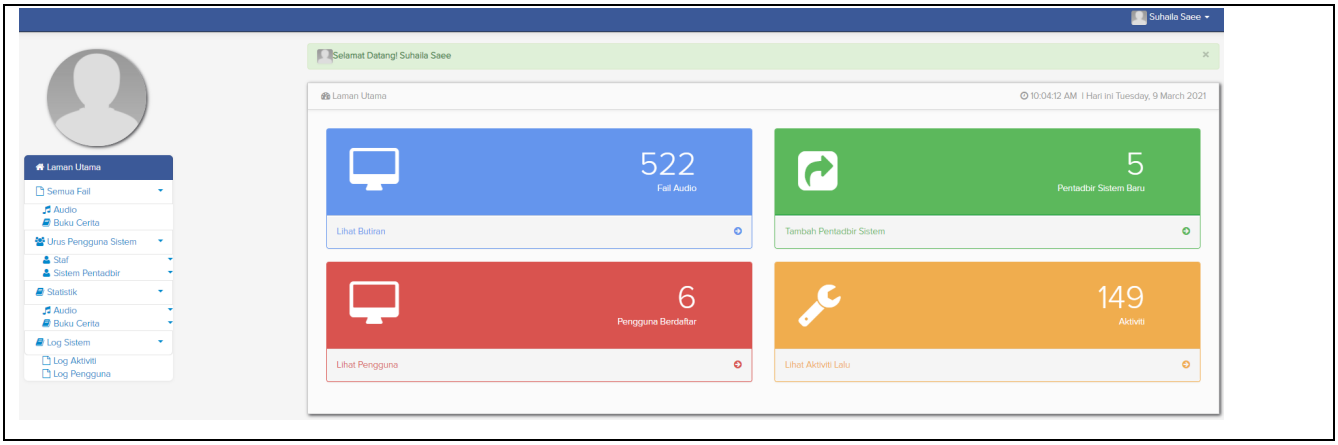
Objektif utama projek ini adalah:

1. Membina sistem berasaskan web (web-based system) untuk menguruskan data digital cerita Borneo, yang akan dipanggil e-CERAB.
2. Menyimpan data digital cerita Borneo yang telah dihasilkan ke dalam sistem e-CERAB.

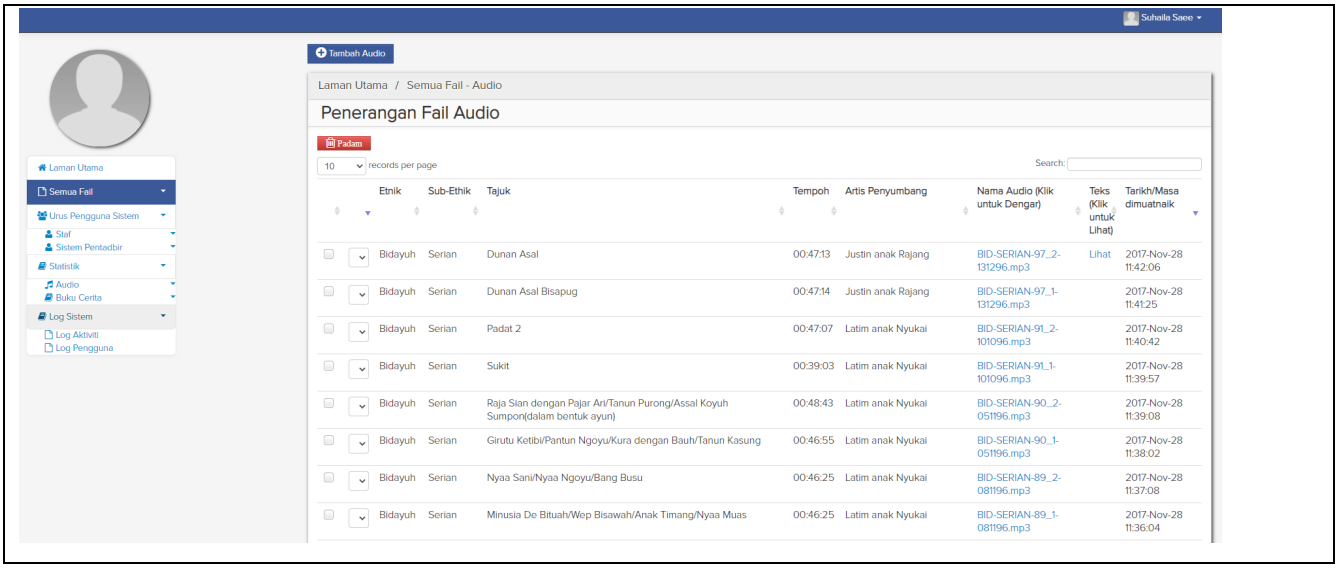
Rajah 1.0 – 3.0 di bawah merujuk kepada antara muka bagi sistem e-CERAB yang boleh diakses oleh pengguna.



Rajah 1.0: Muka depan e-CERAB



Rajah 2.0: Laman utama e-CERAB



Rajah 3.0: Laman muka menunjukkan senarai fail audio dalam format .mp3

Kesimpulan

Pendokumentasian dan pendigitalan bahasa peribumi Sarawak adalah satu usaha yang murni dalam menghidupkan semula dan mengekalkan bahasa pribumi Sarawak. Dengan adanya teknologi digital pasa masa ini, proses pendokumenan dan membangunkan korpora bagi bahasa pribumi Sarawak dapat dijalankan dengan efektif dan efisyen dengan menggunakan kaedah pendokumenan komputasi bahasa. e-CERAB merupakan salah satu contoh usaha dalam mengekalkan bahasa pribumi Sarawak menggunakan kaedah yang dinyatakan. Di harap usaha ini akan berterusan dan dikembangkan lagi dari data audio kepada data teks dalam mengekalkan bahasa pribumi Sarawak.

Penghargaan

Penceramah mengucapkan ribuan terima kasih kepada pihak Dewan Bahasa dan Pustaka Cawangan Sarawak di atas jemputan bagi pembentangan kertas kerja pendokumentasian bahasa pribumi di Sarawak. Pihak Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Malaysia Sarawak juga menghargai kerjasama yang dijalin sejak dari 2017 sehingga sekarang dalam proses pendigitalan kaset cerita rakyat Sarawak dan susulan dari itu, e-CERAB dibangunkan bagi mengukuhkan usaha murni ini.

Rujukan:

Franco, F. M., Hidayati, S., Ghani, B. A. A., & Ranaivo-Malancon, B. (2015). Ethnotaxonomic systems can reflect the vitality status of indigenous languages and traditional knowledge. Indian Journal of Traditional Knowledge, 14(2), 175–182.

Hidayat, S., Ghani, B. A. A., Giridharan, B., Hassan, M. Z., & Franco, F. M. (2018). Using ethnotaxonomy to assess traditional knowledge and language vitality: A case study with the vaie people of Sarawak, Malaysia. Ethnobiology Letters, 9(2), 33–47. https://doi.org/10.14237/ebl.9.2.2018.740

Geiger, D., Rosemann, M., Fielt, E., & Schader, M. (2012). Crowdsourcing information systems-definition typology, and design.

Rehg, K. L., & Campbell, L. (Eds.). (2018). *The Oxford Handbook of Endangered Languages*. Oxford University Press.

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. Journal of Academic Ethics 4:123–155.

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). Language Vitality and Endangerment.

Simons, G. F., & Fennig, C. D. (2018). Ethnologue: Languages of the World (Twenty-fir). Dallas, Texas: SIL International. Retrieved from http://www.ethnologue.com