Presented at the *Canadian Society for Digital Humanities*, June 2017. Congress of the Social Sciences and Humanities Research Council. Ryerson University, Toronto.

Algonquian Dictionaries Project: Common Digital Infrastructure for Lexicography and Language Documentation

Marie-Odile Junker & Delasie Torkornoo

Carleton University

The Algonquian Dictionaries project is a collaboration involving a dozen dictionaries of Algonquian languages (dictionaries.atlas-ling.ca, Junker, 2013). One of its goals is to develop a common digital infrastructure for lexicography and language documentation, that is both reliable and sustainable. We discuss here how we are meeting the technical aspects related to this goal. These include: open source software, multiple levels of access, offline and online formats, API integration, Model-View-Controller architecture, integrated conventions (best practices) and customization.

The Algonquian Dictionaries digital infrastructure is built with mature open source software to provide reliability and sustainability. Such software has a large contributing community, which is easily accessible for online support. It is free to use and has active code updates.

Dictionaries that are part of the project originally existed in various formats: plain text (electronic or printed), tables, text-based databases such as Toolbox, FileMaker exports, FLEX exports. But these proprietary formats did not give us enough flexibility and control, and for some the cost was prohibitive. Some dictionaries continue to be developed in those formats, but we have established standards in database structure and in synchronisation to ensure integration with our infrastructure.

Our approach is based on language documentation, data collection, editing and distribution, as opposed to computational modelling approaches such as Giellatekno, that require to build a computational model of the language within the infrastructure. We do use computational modelling only for expanding search capacities and enhance users' experience (Arppe, Junker & Torkornoo, 2017).

Most of the dictionaries are ongoing projects, with active dictionary teams (Junker et. al, 2017) who need to have access to the data while data collection is on-going. General users also want to access these works in progress. For that, an Agile development methodology is used during

each cycle of the development process. Community input is a key ingredient for shaping the development of interfaces, layouts, types of permission, information buttons, etc.

We want the community users and stakeholders to have access to the existing (sometimes imperfect) and updated data during the lexicographic development. Therefore, multi-levels of access to the data is a core principle of the infrastructure. It required designing the interface around each stakeholder's goals and process. For example, we want certain editors or research assistants to only attach media to existing dictionary entries without being able to modify other content (Figure 1-2).

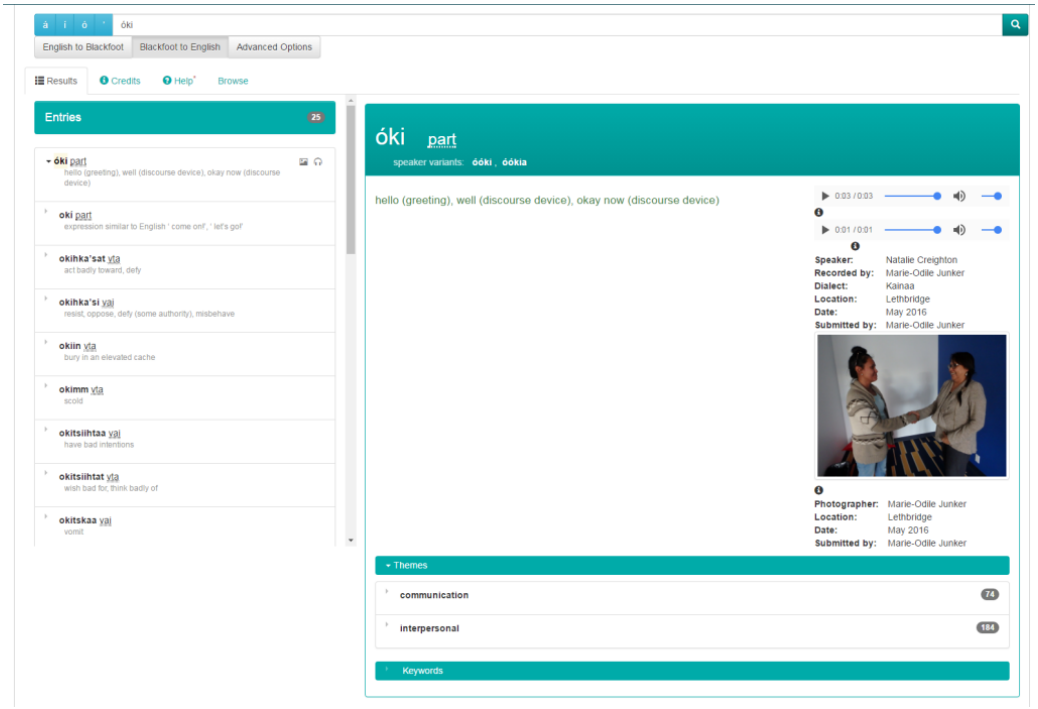

Figure 1: Blackfoot Dictionary general user view with rich metadata displayed

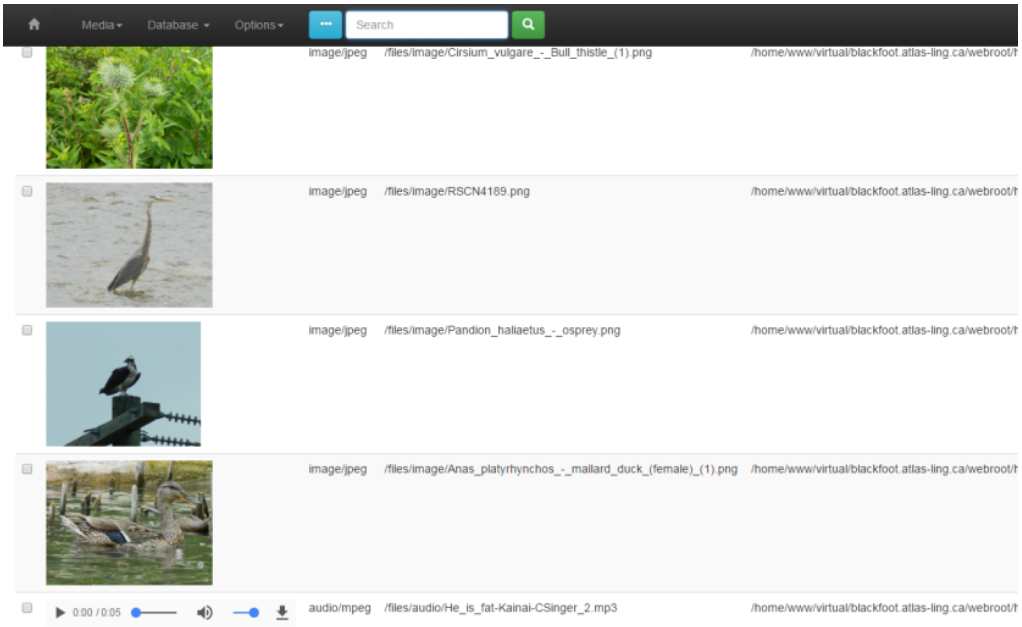

Figure 2: Multimedia editing interface

The infrastructure must be designed around two main categories of end-user, based on their increasing ability to modify the data. Public users do not have modification privileges on the data, but can submit comments and suggestions to the editorial team (Figure 3). Editorial users have varying degrees of data modification ability that is adjusted to their individual evolving capacities. Within a Dictionary editorial team, the editor-in-chief, the general editorial team members, or the data entry team members (research assistants, Indigenous partner organization's staff) must have different levels of permission to modify the data in order to prevent catastrophic data events. Interaction with the system always involves presenting the data to the user based on the level of modification they are allowed to have. The system administrator has to maintain a list of Create, Read, Update & Delete (CRUD) permissions to the data for each user, and their interactions with the data are always piped through their CRUD

permissions. This somewhat personalized and cumbersome administration has proven to be the most efficient for the diversity of users we are involving in such projects.

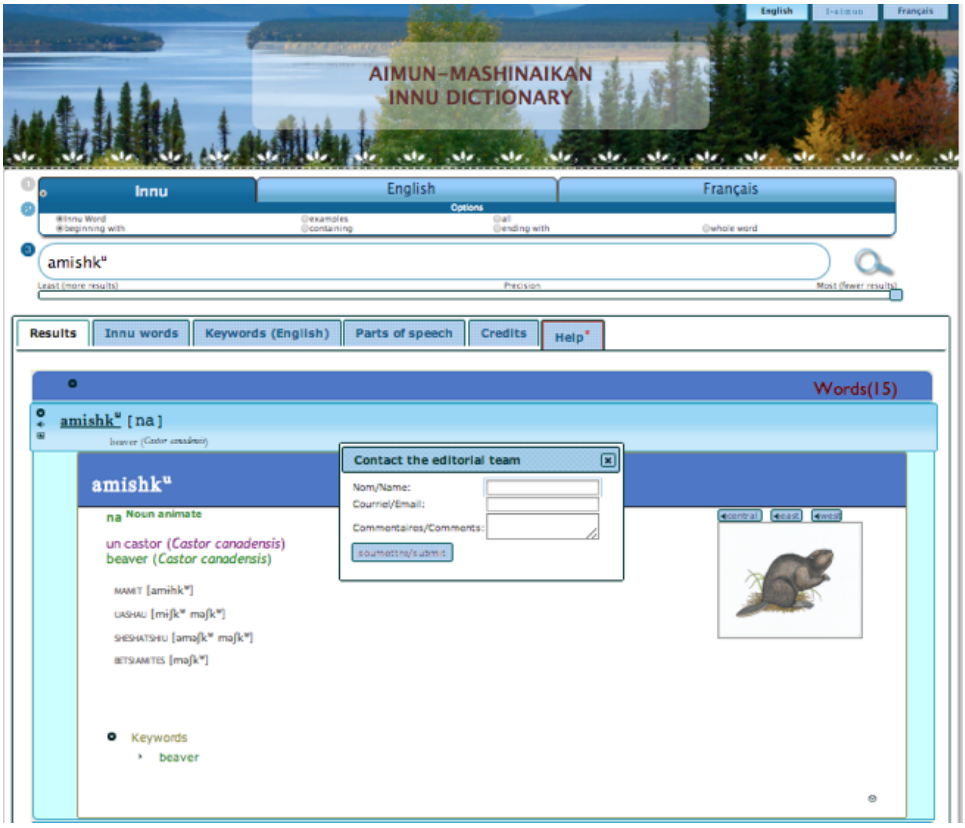

Figure 3: Button for sending comments to editorial team of the Innu dictionary

One of the advantages of a digital infrastructure is the opportunity it gives for multi-format data presentation. The infrastructure can make formats that are accessible online and offline. Online formats have access to live data where changes in the underlining data is instantly available, such as web applications and online mobile apps (Figure 4). Offline formats primarily act as portable formats that are published at prescribed intervals. Our teams have produced printed books, CDs, and offline mobile apps (Figure 5). With the online formats there is the ability to collect precise usage data. For example, in 2017, the Innu online dictionary (Junker&MacKenzie, 2016) had over 100 000 words searched and retrieved. With offline formats we only know the number of downloads or purchases.

Presented at the *Canadian Society for Digital Humanities*, June 2017. Congress of the Social Sciences and Humanities Research Council. Ryerson University, Toronto.

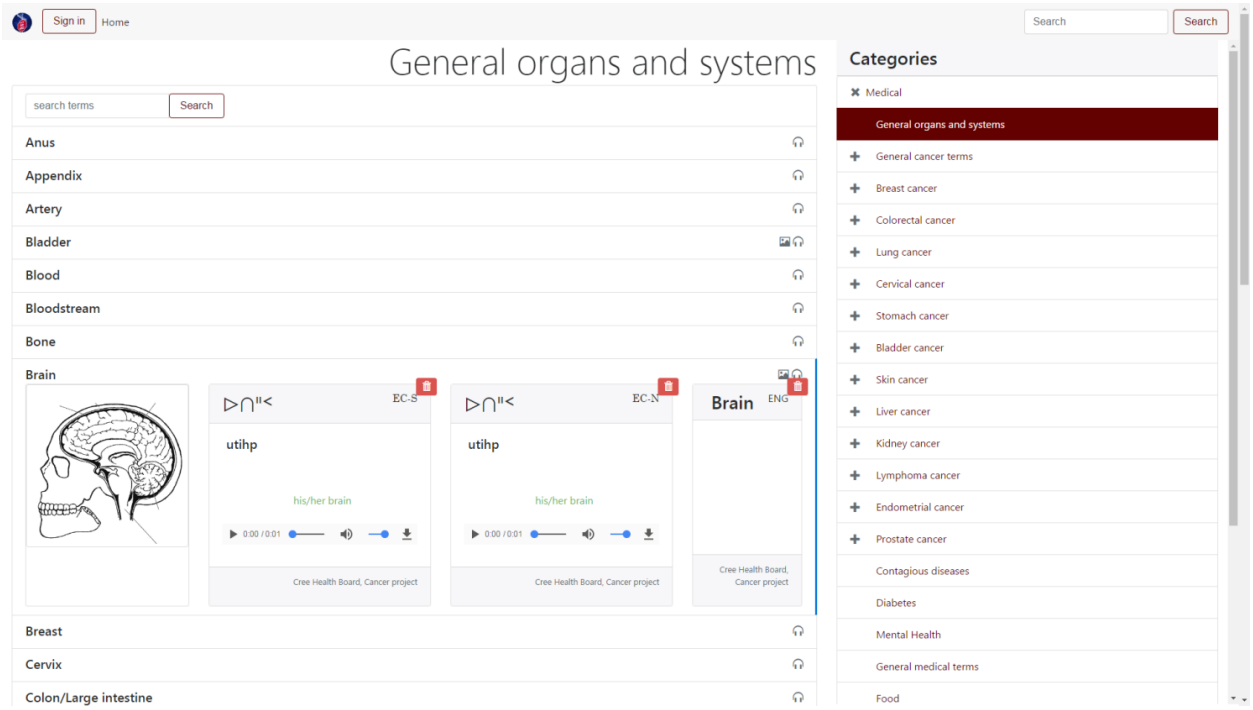

Figure 4: Online format of the East Cree Terminology Forum: terminology.atlas-ling.ca

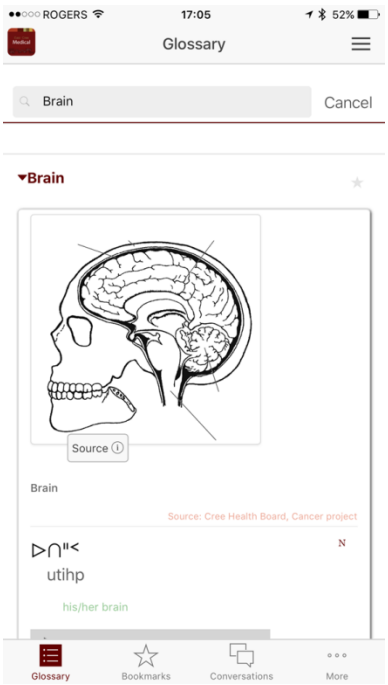

Figure 5: Offline format of the East Cree medical app.

As mentioned above, some dictionaries are still being developed using offline formats, for various reasons, including reliability of internet access. The online formats have been adapted

to accept and synchronize offline development formats. This way we allow individuals to continue to use their favourite mode of preparing the data.

The Algonquian Dictionaries infrastructure is based on a Model-View-Controller architecture. This requires that the data be kept separate from its presentation. A data container is carefully designed for each component of the infrastructure. Using this architecture has allowed the development of multiple views to our data. These views include a web interface (public and administrative view), book (print) format and offline mobile app. With our data (the model), new views can be built by creating a controller that can manipulate the data for the view. Using this architecture has forced us to remove all presentational markup from our data and keep such markups in our views only (for a discussion on how this applies to richly inflected Algonquian verb paradigms, see Arppe et al., 2018). Each view is designed and developed with its own set of presentation markups to communicate data contextually.

MODEL	CONTROLLER	VIEW
Raw information	← Connects→	How users see the information
Source and conceptual representation of the data	The logic to transform data	presentation/formatting of data to your users (admin and public)

Figure 6: An overview of Mode-View-Controller

In order for our dictionaries and language resource websites to expand and share their capabilities, we rely on Application Program Interfaces (API) that communicate with components within and outside of the infrastructure. This allows for functionalities to be built into modules and shared across the infrastructure. For example, the component for handling the general lexicon of a particular dictionary is working seamlessly with its internal multimedia component, and with other external specialized lexicon components, such as WordNet (for English synonyms), or giellatekno (FST). This offers our users the experience of having access to a single large library.

The infrastructure development has required us to develop common conventions and guidelines. The infrastructure now has guidelines on file naming, file formats and folder structures all designed to maximize long term viability. For example, all sound files are kept in a non-compressed format (wav.) for archiving, and a compressed (mp3) version for online use. Sound file and folder names must contain metadata about the speaker and dialect, and a read-me file must be included to describe the recording context. Rich metadata is encouraged in the multimedia manager and sources can be made available when people access the media (see Figure 1 above).

The entire underlining data structure is streamlined for data archiving. There are data protection policies incorporated into the design of the infrastructure to guard again single point of failure events.

Despite the development of common conventions and guidelines, each system is customized according to community and stakeholder feedback and requirements. For example, we noticed that for severely endangered languages, individual contribution sources are preferred, while for more robust languages collective sourcing (to a department, community, workshop event, etc.) is often preferred. We have also customized access, layout, themes, and even data structure.

Conclusion:

We have presented and summarized the technical choices we have made to develop a robust, affordable, and long-lasting digital infrastructure for minority, often endangered and under-funded languages. By working with a dozen languages from the same language family, we are assembling resources and ideas that mutually feed each participating group and enhance the capabilities of this infrastructure. Each time a new language is added, it forces us to develop new solutions which ultimately bring improvements for all participants. While there is a critical mass of participating languages to keep this development going, especially in terms of funding, there is also an ideal size to be found after which levels of customization, exchanges and mutual benefits would diminish. To achieve our long-term sustainability goal, we recommend that technical decisions be systematically based on the use of open source platforms that have a smaller server footprint. The general principles described here should be applicable to other Indigenous or minority language documentation projects.

References:

Arppe, Antti & Harvey, Chris & Junker, Marie-Odile & Valentine, J. Randolph. 2018. Algonquian verb paradigms: A case for systematicity and consistency. In Macauley, Monica & Noodin, Margaret (eds), *Papers of the 47th Algonquian Conference*, 1-22. Buffalo: SUNY Press.

Arppe, Antti, Junker, M.-O. & Torkornoo, D. 2017. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *Proceedings of ComputEL-2: 2nd Workshop on Computational Methods for Endangered Languages.*

Junker, Marie-Odile , Inge Genee, Heather Bliss, Bill Jancewicz, Mary Ann Corbiere, Yvette Mollen. 2017. Building a Common Digital Infrastructure to Sustain Algonquian Languages. Panel presented at the *Canadian Society for Digital Humanities, Congress of the Humanities and Social Sciences*. May 31, 2017, Ryerson University, Toronto. Video available at: https://resources.atlas-ling.ca/video/algonquian-dictionaries-project-presentation-2017.

Junker, Marie-Odile & MacKenzie, Marguerite (eds.). 2016. *Aimun-mashinaikan / Dictionnaire innu / Innu Dictionary.* https://dictionnaire.innu-aimun.ca (accessed 2018-02-20)

Junker, Marie-Odile. 2013. *Summary: A digital infrastructure to sustain Algonquian languages: Dictionaries and Linguistic Atlas*. (Project summary) http://resources.atlas-ling.ca/wp-content/uploads/2015/02/InsightGrant_Algonquian-Project-Summary.pdf.

Websites:

Agile manifesto: http://agilemanifesto.org/principles.html

Algonquian Linguistic Atlas: www.atlas-ling.ca

Algonquian Dictionaries: dictionaries.atlas-ling.ca

Giellatekno: http://giellatekno.uit.no/doc/infra/WhatIsThis.html