ASSOCIATION OF CANADIAN MAP LIBRARIES AND ARCHIVES

BULLETIN

ACMLA NEWS

CARTO 2022 Call for Proposals

The 55th annual conference of the Association of Canadian Map Libraries and Archives (ACMLA-ACACC) will be held virtually from June 13-15, 2022 between 1-3PM EST.

The 55th annual ACMLA/ACACC CARTO conference will be held online. theme this year is Mapping Through a Changing World. The event will include presentations, discussion, and other multimedia work in mapping and geospatial fields for this June event, which promises to be a thrilling new take on the old conference format. This event will also include the ACMLA/ACACC annual general meeting, and some surprise social events!!

Le 55e congrès annuel de CARTO, l'Association des cartothèques et archives cartographiques du Canada (CARTO-ACMLA-ACACC), se tiendra virtuellement du 13 au 15 juin 2022, de 13h à 15h30 HNE.

Le 55e congrès annuel ACMLA/ACACC CARTO se déroulera en ligne. Le thème de cette année est « Cartographier dans un monde en mutation ». L'événement comprendra des présentations, des discussions et d'autres activités multimédias dans les domaines de la cartographie et du géospatial, ce qui promet d'être une nouvelle approche stimulante de l'ancien format de conférence. Cet événement comprendra également l'assemblée générale annuelle de l'ACACC/ACMLA, ainsi que des événements sociaux!

The Call for Proposals is available at:

English Version

French Version

Geospatial Data and Software Reviews

Meg Miller
University of Manitoba

OpenRefine: An Approachable Open Tool to Clean Research Data

Authors
Meg Miller, GIS & Data Visualization Librarian, University of Manitoba: meg.miller@umanitoba.ca
Natalie Vielfaure, Digital Curation Archivist, University of Manitoba: natalie.vielfaure@umanitoba.ca

Abstract
This review provides an overview of data cleaning tools and discusses why and how OpenRefine has been an effective tool in the delivery of one-shot instructional sessions on data cleaning in an academic library context.

Introduction
The following will examine OpenRefine, an open-source Java based program that runs locally. This application can be used to clean and transform data. The tool is very flexible with a modular flow that lends itself well to on-the-fly modification in a classroom context. Here, the tool is used throughout a session to illustrate different data cleaning techniques and considerations as applied to research data and integrating secondary data sources.

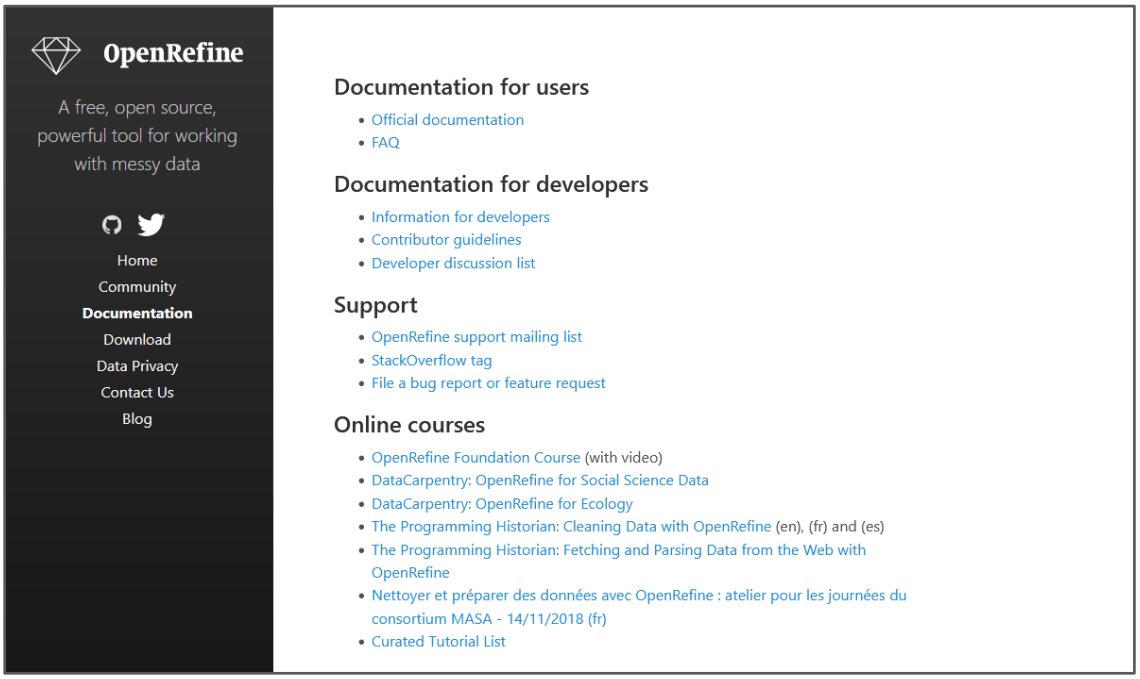

Figure 1: OpenRefine documentation landing page

Session Background

Data cleaning sessions were developed in response to user feedback requesting more sessions covering the steps of the data analytics cycle surrounding visualization. Initially Microsoft Excel was the tool of choice for classroom instruction, but as the service grew to encompass more open tools, it was discovered there was an appetite for OpenRefine outside of one-off consultation style support.

At the beginning of the session, survey data that has yet to be cleaned is presented in a wordcloud to quickly highlight problems in the dataset. Another wordcloud will be generated at the end of the session to help users visualize the impact of the data cleaning actions applied in OpenRefine.

Housekeeping (data management vs. data cleaning), faceting, clustering, merging, splitting, appending and more are all covered in relation to researcher data within the session. OpenRefine also allows you to capture a log of all actions taken on a dataset, which is a requirement for some journals and granting agencies to support a move toward open science. This log can also be used to repeat the actions you take on multiple files.

Tool Details

OpenRefine is an open desktop tool for cleaning messy data. It was previously known as Google Refine. In 2012 Google stopped supporting the project and maintenance has been taken over by a dedicated team of volunteers from around the world (https://github.com/OpenRefine/OpenRefine#credits).

OpenRefine looks like a spreadsheet but operates like a database in a web browser. It allows for a low barrier entry point and a data cleaning alternative to Microsoft Excel. The GUI means that novice users can feel confident in applying basic data cleaning strategies and algorithms, while the advanced functions and use of GREL (General Refine Expression Language) allow for more advanced use cases and scenarios.

The application keeps data private on your own computer and works by running a local server that you interact with via your web browser of choice. The GUI looks like a spreadsheet, but operates like a database, allowing for increased discovery capabilities beyond programs like Microsoft Excel. It comes with many pre-loaded transformations, but expressions can be written in General Refine Expression Language (GREL), in Jython (i.e., Python), and in Clojure. There are many extensions, reconciliation services, and client libraries available to users, as well as robust communities of support that can be tapped into via mailing list or StackOverflow.

To start out, users need to download and install OpenRefine on their workstations. The sample dataset is loaded, and series of actions are taken on the dataset. Major categories are listed below, with ones that lend themselves well to simple modification highlighted.

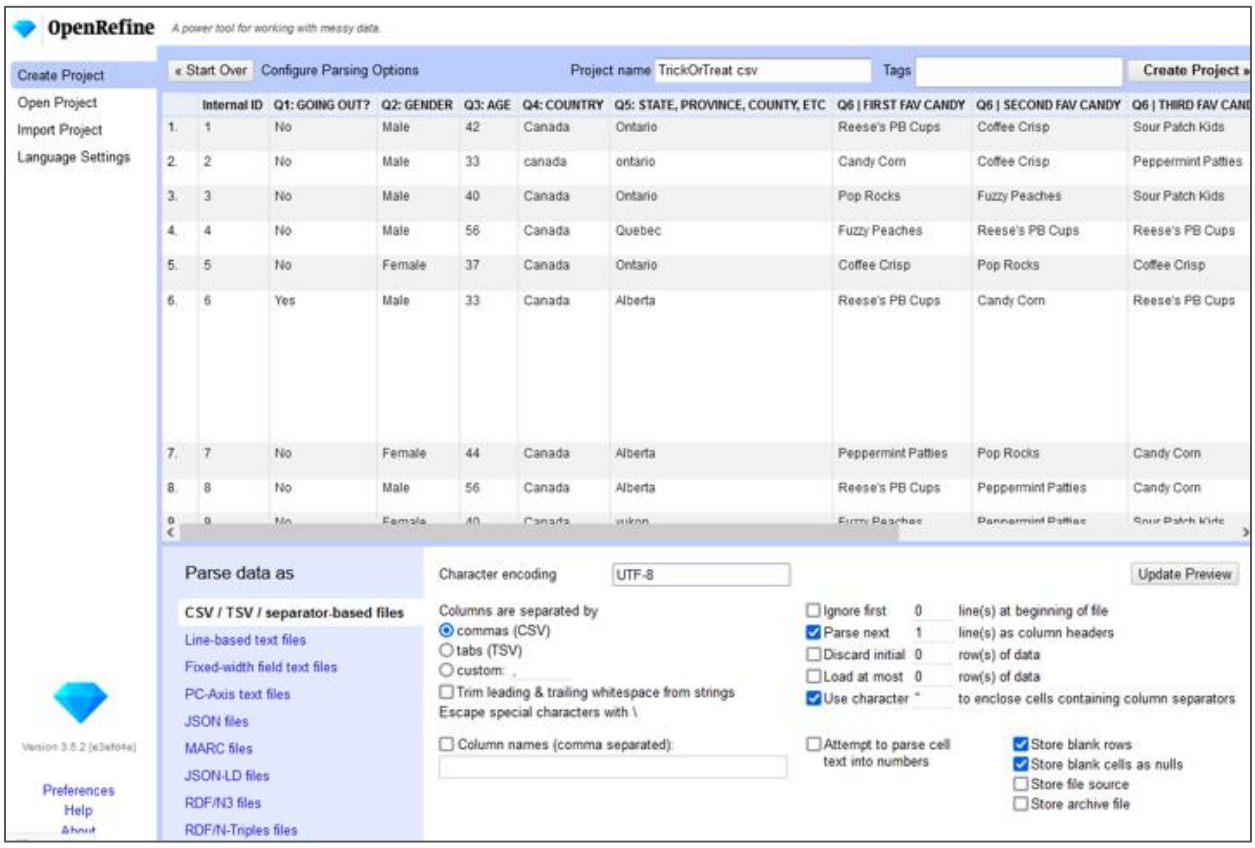

Figure 2: OpenRefine Preview window

Format Support

OpenRefine organizes your data into a project file. When you are transforming the data, it is not the raw data itself that is being manipulated, but the project file. A project file can be created by importing data from your computer, the web, your clipboard, a database, or Google Drive. Many formats are supported, but the most relevant for this audience would be: CSV, TXT, JSON, XML, XLS, MARC and RDF.

Housekeeping

If you supply two or more files for one project, the files' rows will be loaded in the order that you specify, and OpenRefine will create a column at the beginning of the dataset populated with the source URL or file name to help you identify where each row came from. If the files have columns with identical names, the data will load in those columns; if not, the successive files will append all their new columns to the end of the dataset.

If your dataset is more than one million rows you will need to increase your RAM and the server memory allocated. Additionally, the software you are importing from or exporting to may have a different set of limitations to consider.

This is a useful place to discuss data conversions, residency, field types, and the importance of planning one's project.

Dataset structure – Columns

Actions that would be taken on a column include deleting, renaming, adding, hiding, sorting and moving columns. More advanced transformations like splitting and joining columns will be covered in a later section.

This is a good starting point for users to familiarize themselves with navigating the program interface.

Data contents – Cells

Common cell transformations would include trimming whitespace, changing type case, field types and null values. In OpenRefine, changing values to title case is as simple as clicking the dropdown arrow in the header row for the column in question (1), selecting *Edit cells* (2), then *Common transforms* (3), and finally *To titlecase* (4). Other transformations that can be done using GREL are found under the *Transform* option.

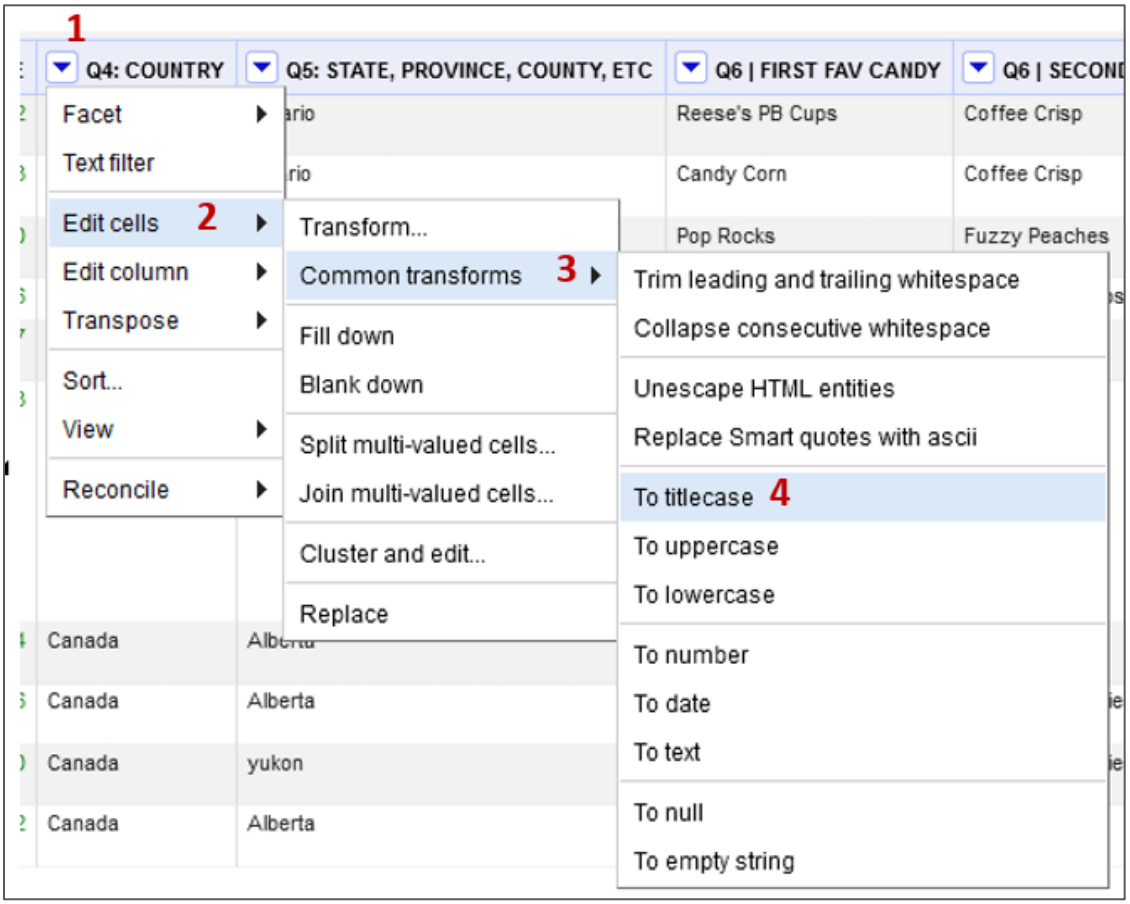

Figure 3: Workflow to change values to title case

Facets

Faceting provides the user with a snapshot of the entries in a particular column and allows them to filter down to a particular record. I use it most to quickly highlight problems with the data and look for outliers. The facet provides a list of cells in a given column to better assess the big picture for that column and further allows you to filter to some subset of rows for which the cells in that column satisfy some constraint.

Edits to the data can be manually made from the Facet pane, where values can be sorted alphabetically (name) or by frequency (count). In an instructional session, this section transitions well into more advanced transformations such as clustering.

Figure 4: Facet view of sample data column

Popular transformations

This section discusses the most popular transformations for which researchers request support. In the session, discussion focusses on not blindly trusting the technology and being an active/considered participant in the data cleaning process.

Cluster

Clustering uses AI and fuzzy matching to best guess which pieces of text refer to the same thing and highlight potential inconsistencies that could be resolved. Opportunity is given to play with the *Method* and *Keying Function* to change the type of algorithm used to fuzzy match. The program does a great job of matching the correct values, but as with anything automated, users are reminded to go through and look at their data for any discrepancies after applying a change. Workflows are discussed at this point: Depending on the data set, what makes the most sense? Running a different clustering algorithm or merging the remaining clusters by hand.

Join

There are many different reasons a user might want to merge or join multiple columns together. Discussion here usually centers around matching field types, and column sizes so records do not get truncated.

Split

Splitting is the most common transformation users request support with, either for parsing text or separating coordinate pairs for plotting data on a map. Working through an example and demonstrating how to use the facets to display records that did not split properly is useful, as you aren't just demonstrating an action, but how to troubleshoot. This example further reminds users of the importance of data cleaning, as consistently structured data will make the splitting process easier.

Extensions

Many plugins have been developed for OpenRefine. The ones that are most relevant to someone in a role supporting data visualization in an academic setting are:

- GeoRefine: Adds GIS functionality to GREL library
- OSM Extractor: Imports OpenStreetMap data using Overpass API
- GeoJSON Export: Exports data as GeoJSON using coordinate pairs or WKT

Note that some operations will involve connecting to an external service. The interface does a decent job at documenting this, but user discretion is advised. A rule of thumb if you are using a service for reconciliation or connecting to an external URL (database, Google Drive, Geocoder...), is read the documentation about how your data is being stored.

In a classroom context, no time is spent actively exploring these plugins. Instead, discussion is centered on seeking help in discussion forums, not reinventing the wheel, and knowing when to step back.

Many examples from each of these actions and more are available on the project site with the code heavily commented to explain what is going on. A link to the documentation can be found at https://docs.openrefine.org/ .

Session workflow (1-1.5hrs)

How does this play out in an instructional session?

Before the session:

1. Students were instructed to download the most stable long-term release of the software and have it installed on their machines before the session.
2. Learning materials (dataset, slides, walkthrough, and FAQs) are uploaded to the GitHub workshop space.

The session:

Topic	***Action***
Session introduction	*Participants are reminded to download and install materials if they have forgotten.*
Overview of Data Cleaning and Software	*Through a slide deck, participants are provided with a definition of data cleaning, common steps that may be applied to a data set to improve its overall quality, and overall benefits of data cleaning for the researcher and secondary users of the data. Participants are then given an introduction to the OpenRefine, as well as Voyant, which will be used to generate a wordcloud.*
Introduce Example Scenario	*Participants are cleaning up survey data to be imported into a GIS software.*
OpenRefine Exercise	
Housekeeping & Importing	*Participants examine the contents of the data folder and preview data using OpenRefine.*

Cell Transformations • To titlecase • To number	*Participants convert:* • *Country column to titlecase* • *Age field from string to number*
Faceting & Clustering	*Participants apply a facet to the Province column to explore the state of the data. Manual editing and outliers are discussed.* *Participants explore the options available in the clustering dialogue and the limitations.*
Column Transformations • Merge • Split • Rename • Delete	*Participants:* • *Merge 3 columns together using a separator* • *Split a coordinate pair column into two columns and trim whitespace* • *Rename the split columns to have meaningful names* • *Delete an empty column (and discuss why hiding can be a better option)*
Advanced Techniques	*Participants discuss their advanced needs, facilitator discusses troubleshooting and where to go for support (documentation, support forums etc).*
Exporting	*Participants export cleaned data as a CSV file and visualize in a wordcloud software such as Voyant or WordArt.com and discuss further steps.*
Wrap-up	*Participants are provided with information on library supports for more specific data cleaning questions or issues. The remaining time is used to take questions from the audience.*

Discussion

In talking to stakeholders, it was found there was a lot of interest on and off campus for sessions on data cleaning. Outside of the typical audience, the Digital Curation Archivist and I attracted a diverse group when the session was offered as a GradSteps session. We have also been invited to run the session as a brown-bag to the Manitoba GIS User Group- a provincial community of practice of GIS users from government, academia, non-profits and private industry.

For the future, intermediate level sessions without canned data will be offered as well as targeted offerings for the archival studies program in the history department.

For archives, OpenRefine offers many opportunities to more effectively manage metadata, accession databases, and archival descriptions. Formats supported by OpenRefine, such as CSV, TXT, and, XML, lend themselves well to this type of work. Cleaning up data exported from a legacy database ahead of ingesting it to new or upgraded databases, flagging inconsistencies in descriptive metadata, or correcting errors in text generated through OCR for digitized records can be resource-intensive tasks. OpenRefine can help reduce headaches in this process by automating some of this work to ensure better long-term access and preservation of holdings.

Overall, OpenRefine's ability to perform some of the heavy lifting for the user makes it worthwhile tool to add to an archivist's, librarian's or researcher's data resource toolkit.

ASSOCIATION OF CANADIAN MAP LIBRARIES AND ARCHIVES
BULLETIN

GIS Trends

Barbara Znamirowski

From the Hip – My Experience with an Interdisciplinary SSHRC Grant

Colleen Beard
Librarian Emeritus, Map, Data and GIS
Brock University, St. Catharines, Ontario

Editor's Introduction

Please welcome as guest author our long-standing colleague and friend Colleen Beard, Librarian Emeritus, Map, Data and GIS, Brock University. Colleen's account of her ongoing research opens our eyes to the fascinating environmental history of the Welland Canal. She shares her experience of involvement with a SSHRC Insight Development Grant-funded project, and explains how knowledge of local environmental history and the varied technical skills of our trade have much to add to academic partnerships. As we redefine the nature of our work and reconsider research library mandates, the trend towards becoming full partners in research is important - a wonderful challenge and opportunity for many of us. Thank you, Colleen for continuing to inspire and lead us! We'll see you at Puddy's Bar & Grill for the next HWCMP talk!

A few years ago, it was a little-known fact that buried on the bank of the Old Welland Canal, in downtown St. Catharines, is a 19th century canal schooner – the James Norris.

It certainly grabbed my interest when Dr. Kimberly Monk, Historical and Maritime Archaeologist, asked if I'd be interested in joining "the crew" to recreate a historic maritime landscape - the Shickluna Shipyard c.1840.

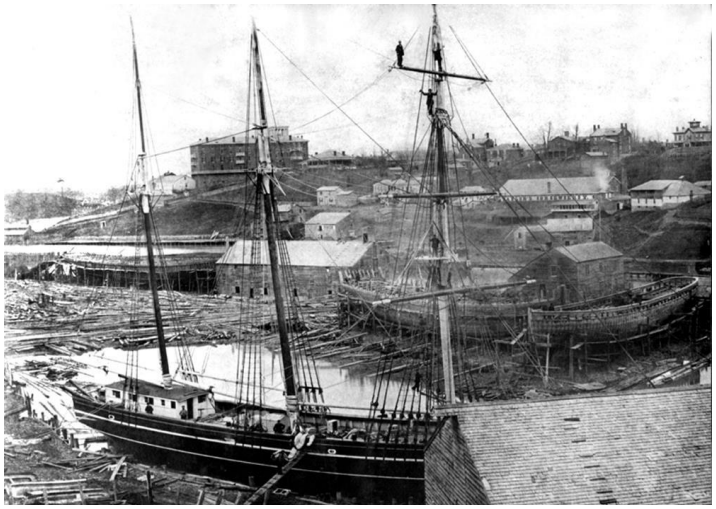

Figure 1. Shickluna Shipyard, 1874. Image courtesy of St. Catharines Historical Museum

Although my knowledge of maritime archaeology is indeed limited, I was sought out to join the team that was successfully awarded a two-year SSHRC Insight Development Grant entitled *Visualizing past landscapes: Toward reengaging the local historic environment*. Yes, it is a rich title, but the goal admittedly is to some day uncover the James Norris c1854 that sits submerged in the now filled basin of the old Shickluna Shipyard - and I want to be a part of it!

If there is one thing I've learned throughout my career it is that the unique skills and knowledge that GIS, map and data librarians possess are undeniably valuable. Perhaps we know this but it is often underestimated. However, not only did the librarian and GIS/data skills seem to be sought out talent for the project, but my expertise of the local canal landscape and the creation of the Historic Welland Canals Mapping Project (HWCMP) was indeed an asset. This is one aspect of my experience I should emphasize. As librarian, part of my job responsibilities was to engage in research. Over many years my studies focused on the three historic Welland Canals that meandered through St. Catharines leaving a legacy of spectacular landmarks. The use of GIS technologies to bring this history alive through interactive visualizations has gained much attention and recognition with the general public, the Brock community, and beyond. Many of my librarian colleagues have applied their GIS skills into similar creative and impressive HGIS research (Fortin & Bonnell, 2014). These skills are unique and often seen as difficult to hone by a typical researcher. But because they can be applied to any discipline, coupled with a bit of creativity, we can contribute greatly to any grant project.

The opening of the Welland Canal in 1829 opened a new route for exports of key staples such as timber and agriculture and led to the establishment of a maritime community which supported the demands of a growing region. It was the main economic activity in Niagara in the 1800's and shaped many communities. The establishment of shipyards was critical to enabling export of high-volume, low- value bulk cargoes on the Great Lakes-Atlantic route and was central to economic development. However, much of the shipyard history is now masked by the built landscape. But the Shickluna Shipyard remains largely abandoned and the filled basin, undisturbed. Although boathouses and the yard buildings no longer exist, potential for archaeological study is high.

In the summer of 2019, as part of the grant requirements, Dr. Monk ran a credit archaeology field course out of the Department of History, Brock University, using the Shickluna site. But this did not come easy. Although we were fortunate that the shipyard sits on city-owned property, the hurdles we had to circumvent were many. Permissions to access the site, insurance, educating city politicians (most unaware of what lurks below on their property) with reports and presentations, etc., is credited to Dr. Monk's tireless efforts and determination. Meanwhile, my first task was to help establish the best location to "dig". Assisting Dr. Joe Boyce, McMaster University, with GPR (ground penetrating radar) reconnaissance of the entire site, the marine environment was not ideal for GPR producing little evidence of submerged remains. Relying on resources from the Welland canals mapping project, I created the Shickluna Mapping Project.

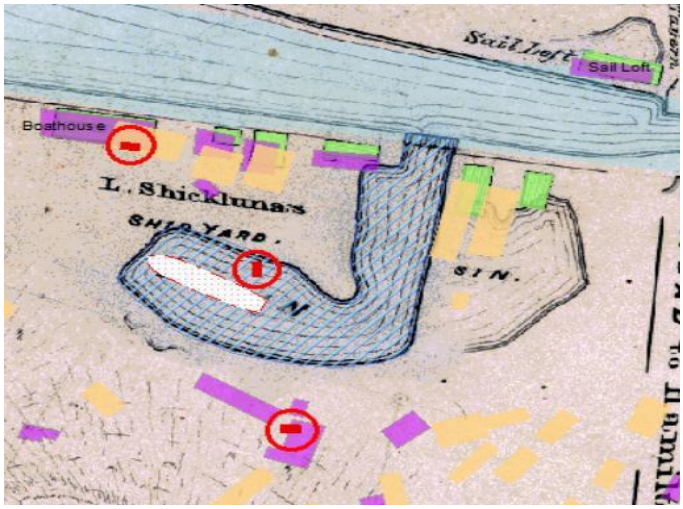

Figure 2. Shickluna Shipyard Mapping Project. Location of Operational Areas (dig sites). Image courtesy of Colleen Beard

ISSN 2561-2263

Based on information gathered from historical maps and photos, a GIS overlay was created to show historical buildings within the shipyard, and the location of the basin with the buried James Norris. Using field techniques such as ArcGIS Collector, we were able to establish three pits, one adjacent to the boathouse; one near the ship; and the other adjacent to the workers' living quarters. This resulted in uncovering over 3000 artifacts – from toilet seats to gin bottles to ceramic figurines – some would describe it as 19th century garbage. But lovely garbage, it is!

Figures 3 and 4. Samples of uncovered artifacts. Photos courtesy of Kimberly Monk

As part of the course curriculum, I gave a talk on the mapping project including the detailed process for locating the "dig" pits, or operational areas. I also led the students on a canal tour as an orientation for "placing" the shipyard in historic context. Throughout the summer the project's progress was documented by social media, such as Twitter and Facebook. The University's *Brock News* was especially instrumental in raising the profile of the project through news features. Later that summer we ran open-house weekends for the general public and local dignitaries to tour the Shipyard; view some of the artifacts; to learn the Louis Shickluna story and the impact his shipyard had on the local community, and sell t-shirts. Using GIS field techniques, we chalked the outline of the submerged ship. The event generated a frenzy of interest from the local community, the local media, and gained an extensive network of *Friends of Shickluna*, including a distant relative of Shickluna. We meet regularly at the historic *Mansion House* c1800 pub (once a store of Hamilton Merritt – the founder of the Welland Canal). More importantly this sparked numerous news articles, radio interviews and greater attention from local politicians and the business community – potential funders for the larger picture! https://www.facebook.com/shicklunashipyard/

Figure 5. Shickluna Shipyard, 2019. Outline of the location of the buried James Norris schooner. Photo courtesy of Colleen Beard

ISSN 2561-2263

This SSHRC grant is a two-year project which aims to: "Develop a framework for reengaging local historic environments; to devise an approach to geo-visualize archival, archaeological, geophysical and environmental data within 3-D model simulations; create tree chronologies for the Niagara Region, toward interpreting historical and industrial impacts; retrace 19th century transport zones within the Great Lakes; generate software to create Narrative Objects, and multi-modal narratives expressed in Virtual Reality and Augmented Reality." (excerpt from the grant submission, 2018). It is important to realize that this is an Insight Development Grant – an ideal steppingstone for subsequent full grants and funding to realize the "ultimate goal" - unearthing the James Norris.

The diverse line-up of grant co-applicants is a great example of building on the expertise of other disciplines working towards a common goal. It seems this is an extremely appealing element with successful SSHRC grants. Leading the project is emerging scholar Dr. Kimberly Monk (History, Brock University), who is a trained maritime historian and field archaeologist, interested in the connections between historic environments, the nature of maritime trade and warfare, and the design of ships. This project builds upon previous research undertaken for her Master's Thesis involving the Welland Sailing Canaller *Sligo*, built by Louis Shickluna in 1860 at his St. Catharines shipyard. Established scholars include: Dr. John Bonnett (History, Brock University), an expert in the digital humanities, specializing in the application of 3-D modelling, including the emerging mediums of Virtual Reality (VR) and Augmented Reality (AR); Dr. Michael Pisaric (Geography, Brock University) who examines the role of paleoecological indicators, toward understanding climate change; and Dr. Joe Boyce (Earth Sciences, McMaster) employs geophysics and geo-archaeological methods to document paleoenvironmental changes in lakes and coastlines. With my expertise in GIS and the Welland Canals, I lead the geo- visualization aims of the project, integrating archaeological, historical, geophysical and environmental data using geospatial mapping technologies. This basically means mapping any data that is generated by all partners of the project, in coordination with the co-applicants.

As a librarian, I also emphasized the importance of including data management practices in the SSHRC Grant application - such as those supported by the Alliance's Portage Network. My role also includes the design of a Data Management Plan (in progress); implementing metadata standards; data storage; and all that stuff that comes with data sharing. Often overlooked, this inclusion is so very important in today's competitive climate of successful grant funding.

Over the past several years I have given over 20 canal talks to various groups, including ACMLA. The HWCMP is technical, but I've become a storyteller in attempts to relate to the not-so-tech- savvy citizen. If there is any advice I can offer it's to limit the academic content and focus on entertaining your audience. Stories of how I encountered coyotes on my canal hikes; discovering other shipwrecks on the canal using historic air photo overlays; Dynamite Luke's attempt to blow up Lock 24 in 1920; and marveled at scenes like this one (*Figure 6*), and how I got inside lock 12 chamber to do 360 filming. More importantly, how I got out!

Our personal knowledge of local history and the use of GIS to share it can be very powerful. Creating visualizations that engage the public is extremely rewarding and takes very little to impress, I've learned. But one must keep in mind that user experience is critical. Ease of use and simple to navigate is key.

Figure 6. 360 filming in Lock 12 chamber, Third Welland Canal. Photo courtesy of Rene Ressler

I became involved with this project during my last "working year" followed by a sabbatical. It was an early retirement incentive. It resulted in a very fluid transition rather than a "what do I do in retirement?" conundrum. As Emeritus status, my research continues with all the resource perks the university has to offer.

The momentum gained in the first year of the project in 2019 was, of course, placed on pause because of the pandemic. However, recent permissions from Brock, the city, and the province will mobilize the project again this summer 2022 with another field school offering. And what was a little-known fact about the buried James Norris a few years ago, has morphed into a full-scale community embraced mission.

GIS Trends: Note from the Editor

Submissions and Feedback

GIS Trends is a place to share ideas, observations and discoveries in the area of GIS and other spatial technologies. If you have something you would like to share please write to me. We also welcome feedback on GIS Trends articles. Proposals for articles and feedback should be sent to: bznamirowski@trentu.ca Thanks for reading and contributing! Barbara Znamirowski, Editor, GIS Trends