Atlanta University Center
Robert W. Woodruff Library

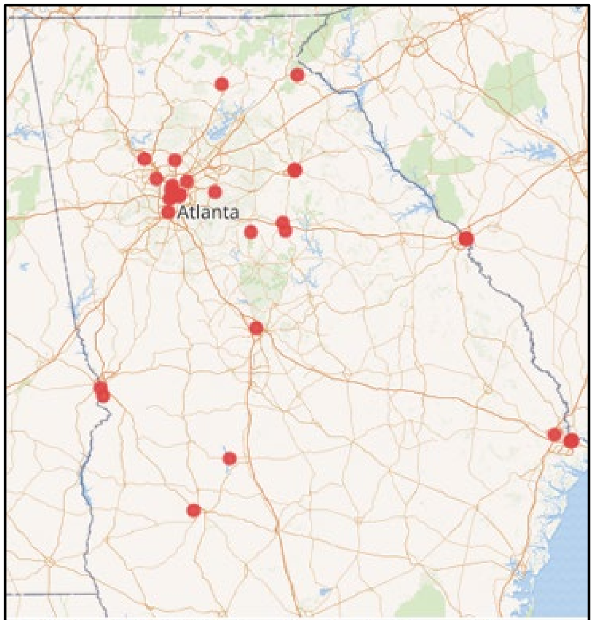
GaNCH: Using Linked Open Data for Georgia’s Natural, Cultural and Historic Organizations’ Disaster Response

This project was made possible in part by a 2019 award from the Catalyst Fund at LYRASIS.

Established in 1982, the Atlanta University Center (AUC) Robert W. Woodruff Library serves the nation's largest consortium of historically black colleges and universities, which includes Clark Atlanta University, the Interdenominational Theological Center, Morehouse College and Spelman College. In addition to the aesthetic benefits of this state-of-the-art facility, the Library has evolved into a model repository of information resources and a front-runner in the innovative delivery of digital resources. The AUC Woodruff Library is the winner of the 2016 Excellence in Academic Libraries Award in the university category from the Association of Collegiate and Research Libraries (ACRL). The AUC Woodruff Library is also home to the Archives Research Center, which is noted for its extensive and unique holdings of materials on the African American experience, including the John Henrik Clarke Africana and African American Collection and the Henry P. Slaughter and Countee Cullen Memorial collections. In addition, at the request of Morehouse College, the Library serves as custodian of the Morehouse College Martin Luther King Jr. Collection.

1. The goal of your project – what did you hope to achieve?

In June 2019, the Atlanta University Center Robert W. Woodruff Library received a LYRASIS Catalyst Fund grant to support the project, "Using Linked Open Data for Georgia's Natural, Cultural and Historic Organizations' Disaster Response." This one-year project's goal was to create a publicly editable directory of Georgia's Natural, Cultural and Historical Organizations (NCHs), allowing for quick retrieval of location and contact information for disaster response. By the end of the project, over 1,900 entries for NCH organizations in Georgia were compiled, updated, and uploaded to Wikidata, the linked open data database from the Wikimedia Foundation. These entries include directory contact information and GIS coordinates that appear on a map presented on the GaNCH project website (https://ganch.auctr.edu/), allowing emergency responders to quickly search for NCHs by region and county in the event of a disaster.



(Wikidata in 2018) Galleries, Libraries, Archives, and Museums (GLAM) in Georgia with coordinate information: 40 institutions

As the largest state east of the Mississippi River with 159 counties, Georgia has extensive and diverse natural, cultural and historic sites that preserve and document the unique history and culture of the state. Including libraries, museums, archives, historical societies, historic sites, state parks, and performing arts organizations, these Natural, Cultural, and Historical (NCH) organizations' sites are critical to cultural tourism. The most recent slate of natural disasters -- including Hurricanes Matthew, Irma, Michael, and Dorian as well as regional flooding, and a very active tornado season -- has resulted in damage to historic structures, important documentary and artifact collections and valued cultural resources. These natural disasters have also significantly impacted – sometimes closed – community organizations important to citizens coping with disasters, such as public libraries and local archives. A rapid response is critical to successfully salvaging NCH organizations and collections that are important to our citizens and economy.

Georgia, however, did not have a means for quickly identifying NCH sites in areas impacted by disasters, thus impeding response. Having an up-to-date and comprehensive inventory of existing NCH sites is essential to preventing loss of important NCH resources and limiting the financial impact of a disaster. Only with accurate geographic locations of these resources will emergency responders and cultural heritage managers be able to work together to expedite the recovery process. The Georgia Emergency Management & Homeland Security Agency (GEMA) will integrate this database in their web-based Emergency Operations Command software (WebEOC) to identify locations that have been impacted during an event and in facilitating response recommendations.

This project's approach leverages linked open data to not only solve an existing problem for NCHs in Georgia, but also provides a model for other states. Project benefits include:

1. Flexible structure – Wikidata is able to represent many relationships in repeatable fields; records are easily enhanced with additional data
2. Decentralized updates – Simple edits can be done manually by anyone; no need for an institution or individual "owner" to grant access to the data for regular updates, thus removing bottlenecks
3. Open license – Facts represented in Wikidata are public domain
4. Sourced information – Reference links are included for each statement, for verification and to provide direction for future updates
5. Broad access and free implementation – The project utilizes free software
6. Model for other states – By documenting and presenting on the processes used, the project provides other states with a model that can be easily reproduced
7. Graduate student paid internship – This project provides a graduate student in Library and Information Science, Archival Studies or related field with practical experience working with linked open data.

2. Your model or process

In 2018, the Atlanta University Center Robert W. Woodruff Library (AUC RWWL) conducted a pilot project to test the feasibility of updating an existing but outdated NCH data set. The Georgia Historical Records Advisory Council's *Directory of Historical and Cultural Organizations* was web-scraped and placed into a table. The data was enhanced using free web tools to include geospatial coordinates and county information. A subset of ten records was reconciled against Wikidata using OpenRefine, a free data cleaning tool. This subset was then uploaded to Wikidata, enabling us to query the data and display the results using Wikidata's Query tool. Two example displays were created and published online using the Wikidata Query Service (http://clifflandis.net/WikiData_GA-CHO-DR_test.html). These examples showed 1) an auto-generated list of contact information for NCHs in an area (City/County/State), and 2) an auto-generated map of all NCHs in a geographic area.

The project team consisted of staff members at the Atlanta University Center Robert W. Woodruff Library: Cliff Landis, Digital Initiatives Librarian, (Co-PI); Christine Wiseman, Department Head, Digital Services Department (Co-PI); Allyson F. Smith, Graduate Student Assistant; Matthew Stephens, Web Developer; Jessica Leming, Digital Curation Librarian; and Alex Dade, Library Technical Specialist. The project operated under the leadership of Loretta Parham, CEO/Library Director.

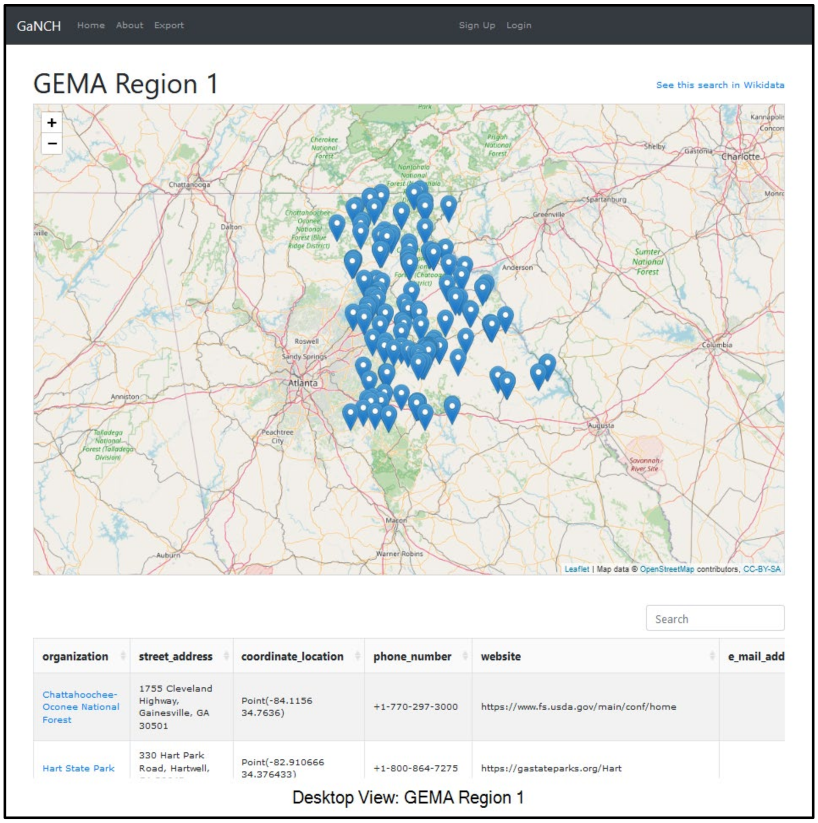
a. What did you do

The success of a state-level project like this would not be possible without the support from partner organizations. Early in the project, we reached out to cultural heritage emergency response organizations who would be using the resulting datasets and website: GEMA, HERA (Heritage Emergency Response Alliance), and SHER (Savannah Heritage Emergency Response). We also reached out to government organizations and professional associations to advertise the project and get source datasets. After we got the initial support from our project partners during the grant application project, we set up bi-monthly meetings with the partners to get feedback as the project progressed, as well as to show the work as it developed. These project partner meetings were recorded with Zoom and uploaded to the Internet Archive as a way to show how the project progressed over the course of the year. A full list of project partners can be found on the project's GitHub site (https://github.com/clifflandis/GaNCH/blob/master/docs/project_partners.md).

After the successful pilot workflow test, we scaled-up the workflow into a seven-step process:

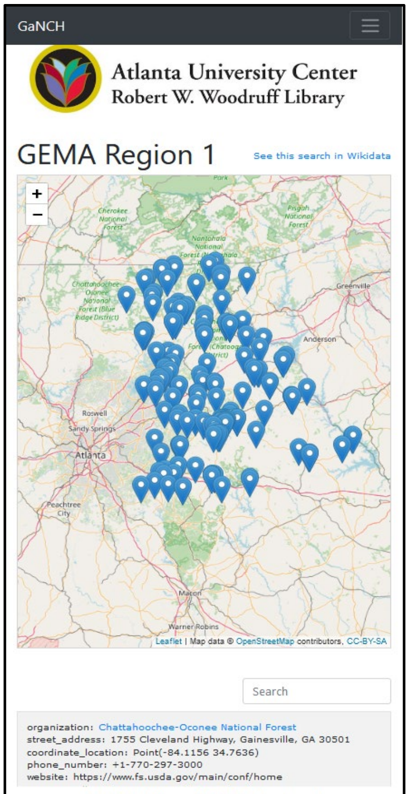
1. Receive dataset - use Python scripts to scrape websites, download available datasets, or request datasets from project partners
2. Format to spreadsheet template - copy relevant data over to the dataset template in CSV (comma separated value) format
3. Index and remove duplicate entries - search each organization name in Wikidata to see if a record exists, and to see if we've already created/edited the record in a previous upload. If so, delete the duplicate entry to prevent performing duplicate enrichment work
4. Gather, verify, enrich, and create references for data - perform web, phone, and email research to verify the existence and directory information for the organization. Enrich with county and geocoordinate location information. Use the Internet Archive Wayback Machine to create snapshots for references.
5. Reconcile in OpenRefine - we created a data dictionary to identify and prioritize which data points were most valuable for disaster recovery, and to match them to Wikidata's data model (https://github.com/clifflandis/GaNCH/blob/master/data/data_dictionary.md)
6. Upload dataset to Wikidata - upload the revised and referenced records in the dataset to Wikidata.
7. Perform quality control on ingests - review each record in Wikidata and make any post-uploads necessary. Examples include verifying geocoordinate locations, removing duplicate field entries, and adding parent/subsidiary organization relationships.

To help other organizations who may want to adapt or reuse this workflow, we've developed a Workflow Manual that walks step-by-step through this process in greater detail (https://github.com/clifflandis/GaNCH/blob/master/docs/workflow.md).


GaNCH
Home
About
Export
Sign Up
Login
GEMA Region 1
See this search in Wikidata
+
−
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA
Search

organization	street_address	coordinate_location	phone_number	website	e_mail_add
Chattahoochee-Oconee National Forest	1755 Cleveland Highway, Gainesville, GA 30501	Point(-84.1156 34.7636)	+1-770-297-3000	https://www.fs.usda.gov/main/conf/home	
Hart State Park	330 Hart Park Road, Hartwell,	Point(-82.910666 34.376433)	+1-800-864-7275	https://gastateparks.org/Hart	

Desktop View: GEMA Region 1


GaNCH
Atlanta University Center
Robert W. Woodruff Library
GEMA Region 1
See this search in Wikidata
+
−
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA
Search

organization: Chattahoochee-Oconee National Forest
street_address: 1755 Cleveland Highway, Gainesville, GA 30501
coordinate_location: Point(-84.1156 34.7636)
phone_number: +1-770-297-3000
website: https://www.fs.usda.gov/main/conf/home

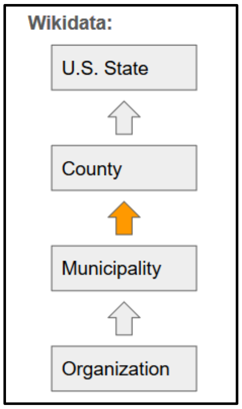
Mobile View: GEMA Region 1

In addition to the data workflow, we developed a mobile-friendly website that can use prepared linked data queries to display both a map of NCHs and a table of their contact information. In April 2020, we facilitated a focus group with emergency response coordinators at HERA, GEMA, and SHER to ensure that the final website will meet cultural heritage emergency responders' needs. After gathering that feedback, we have developed three views for users: 1) All of Georgia, 2) Georgia Counties, and 3) GEMA Regions. The website also provides instructions for users on how to export search results via the Wikidata Query Service. This enables cultural heritage emergency responders to integrate the results into the state-level WebEOC system.

b. What worked and what did not

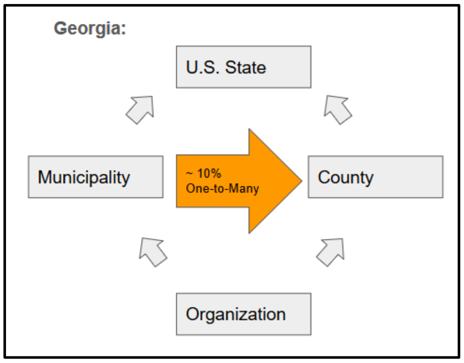
Overall the process worked well. The most frequent concern that people have raised about using Wikidata for this project is the risk of vandalism, but over the last year we've only seen four minor instances of unhelpful edits, which were easy to revert. It was a pleasant surprise to see that the most commonly raised concern did not present a challenge for the project.

However, we did encounter a few challenges. One problem we have encountered is outdated information persisting in the initial datasets we received-- in some cases NCH organizations had been closed for over a decade, yet still persisted in government and professional organization directories. When possible, we included these dissolved organizations in our uploads, and included the dissolution date with a reference to show evidence that the organization no longer exists. This not only provides public documentation of the organization's closure, but it also allows us to exclude these closed organizations from our search results to prevent cultural heritage emergency responders from wasting time during a disaster.



Another issue we discovered is the challenge of trying to map real-world messiness to the much cleaner data model of Wikidata. The challenge was not only trying to map it, but trying to reach consensus with the Wikidata community on how it should be handled.

In Wikidata, the P131 field is the field for administrative territorial entities in Wikidata. This P131 field is broad and can cover municipalities, counties, and states. In the Wikidata documentation for P131, it recommends that you only list the *single most local* administrative territorial entity, since the field is supposed to be both transitive and hierarchical, and cascade upwards like you see in the diagram on the left.



But in Georgia, the borders of municipalities and counties were drawn independently, and about 10% of the municipalities in Georgia exist in more than one county, so they are not transitive. This means that our project cannot rely on Wikidata's cascading hierarchy to be correct when it comes to searching for organizations by county.

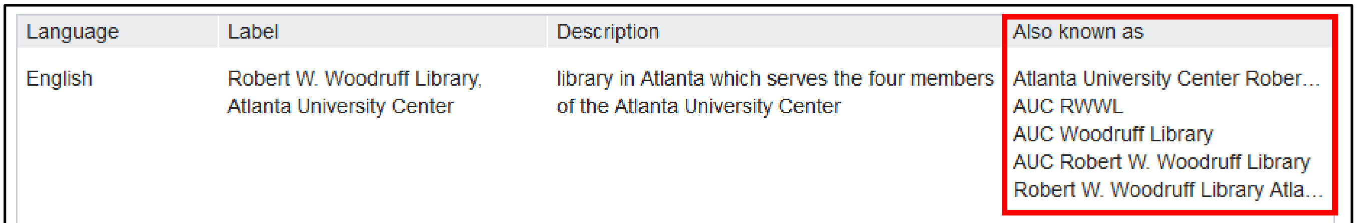
We reached out to the Wikidata community to try and find a solution to this challenge. After several proposals and over a month of discussion on four separate Wikidata pages, no consensus was reached. As a result, we decided to explicitly declare municipality, county, *and* state, all in the P131 field, with the hope that consensus on a solution can be reached in the future. In the meantime, our queries are functioning well, and we figure it's better to have too much well-sourced information rather than too little.

To help other organizations who may adapt or replicate this project, we have documented the challenges that we encountered and how we addressed them (https://github.com/clifflandis/GaNCH/blob/master/docs/challenges.md).

c. What modifications did you make from the original proposal, or would you recommend others make if they want to adapt your model?

There were a few parts of our initial plan that had to be modified to ensure high quality results for the project. First, we initially planned to use free, online geocoding tools to obtain geocoordinate locations for the NCH organizations. Unfortunately, since these tools rely solely on the address of the organization, they can often yield incorrect results. After a disaster, wayfinding markers, or even whole buildings can be destroyed; it was important for us to identify the physical locations of these NCH organizations, so the mailing address or nearest intersection was often insufficient. As a result, we shifted to manually locating each NCH organization's geolocation coordinates via Google Maps, which in some cases required a fair amount of research. For example, most of the technical colleges in Georgia have multiple campuses, each with their own libraries. Many of these libraries do not have their own webpages to reference, they often exist in shared facilities (e.g. Academic Building C), and they are not distinctly identified in Google Maps. By looking for online campus maps and cross-referencing with Google Maps, we were able to identify the physical location of each library.

We also encountered a workflow challenge with NCH organizations' name variations. We knew in advance that many of the organizations would appear in multiple source datasets, therefore we created an index file to assign accession numbers for each organization. As we prepared new datasets, we compared the organization names against the index to prevent duplication of our work. We also knew that organizations' names would vary between datasets, so we planned to check each dataset against the index using the Microsoft Excel Fuzzy Lookup Add-on to identify similar/identical organization names. Unfortunately, this method was insufficient at identifying duplicate records, therefore we changed our workflow to check the names of organizations manually in Wikidata before enriching the data. As we discovered new name variations, we added them to Wikidata as Labels (alternative names). This improved our results by reducing duplication of work, however, a few duplicate records of organizations were present on each dataset that had to be cleaned up manually during quality control in Wikidata. A bonus to this process is that we generated a clear list of how many organizational records were edited (via the index), and that Wikidata now includes a comprehensive list of all the name variations we encountered for each organization.



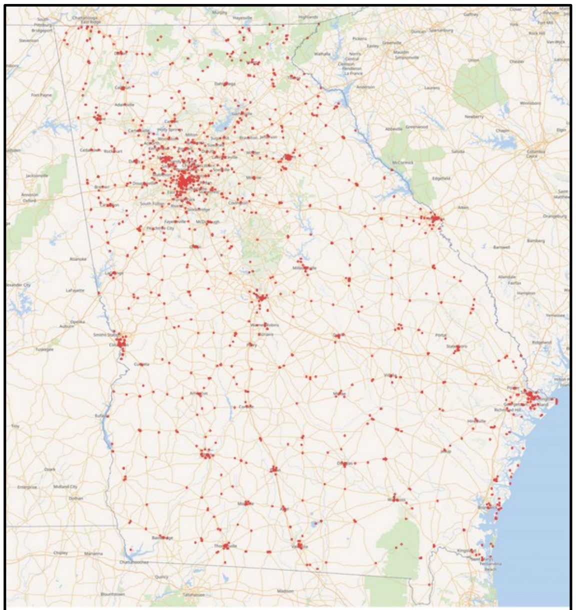
Language	Label	Description	Also known as
English	Robert W. Woodruff Library, Atlanta University Center	library in Atlanta which serves the four members of the Atlanta University Center	Atlanta University Center Rober… AUC RWWL AUC Woodruff Library AUC Robert W. Woodruff Library Robert W. Woodruff Library Atla…

Recording Labels (name variations) in Wikidata

We would recommend approaching a project like this with patience and flexibility. The COVID-19 pandemic struck when we were about 60% through the project, which required us to transition to remote work. Thankfully, none of the work associated with this project required staff to be physically located in the library, and we were able to successfully transition to working from home. However, we recommend having "backup tasks" ready when free online tools are temporarily slow or unavailable. On several occasions throughout the project there were days where either Wikidata or the Internet Archive Wayback Machine would temporarily be slow or unavailable. On days like these, we switched to work which did not require using the slow/unavailable tools. Patience and flexibility are vital when roadblocks require changing tasks.

Lastly, if we were to repeat this project we recommend asking for additional funding in two areas: to pay for promotional activities, and to pay for an additional graduate student assistant. We did not originally allocate any funds for travel costs to spread the word about the project, therefore our opportunities in this area were limited. However, in the wake of COVID-19, many conferences transitioned to a completely virtual format and waived their registration fees for presenters. This provided us with additional opportunities to present on GaNCH at the state and national level. Although we were able to accomplish a great deal, with additional funding for a second graduate student assistant we could have expanded the project to include additional NCH organizations (such as county Clerk of Court offices). We also could have further enriched the records for the organizations that were included. With limited time we had to cut back on recording social media accounts, inception dates, and parent organization/subsidiary organization relationships part way through the project to ensure we were able to capture core data for more organizations.

3. Your accomplishments – what did you achieve?



(Wikidata in 2020) NCH organizations in Georgia with coordinate information: 1960 institutions

At the beginning of this project, there was no easy way for cultural heritage emergency responders to reach out to NCH organizations before or after a disaster. The data available was scattered across the internet in disparate databases and tables, and was rarely up-to-date. What started in 2018 as 40 GLAM institutions mapped in Wikidata has now grown to 1,916 NCH organizations indexed as of August 2020.

We retrieved, verified, updated, enriched, and uploaded thirteen datasets gathered from our partner organizations, each with as few as 15 to as many as 556 records.

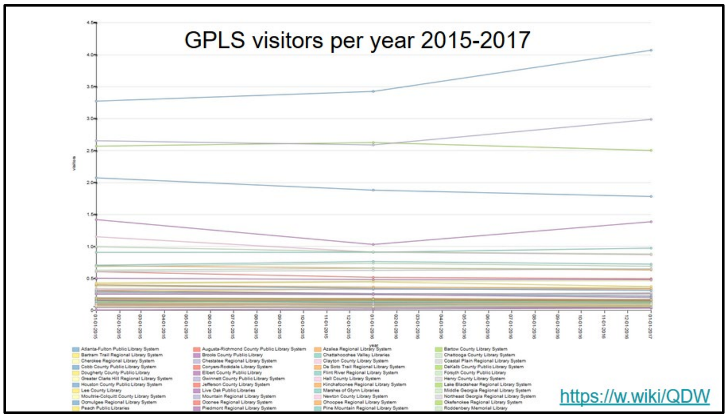
Additionally, we have broadened the scope of our search queries to meet the needs of cultural heritage emergency responders. For example, our search queries now include historic districts which, although they don't have contact information, are important for GEMA fly-overs to assess damage of historic properties not associated with cultural heritage organizations. As a result, our largest search query now includes 1,960 results.

Our website will serve as a one-stop location for Georgia's cultural heritage emergency responders to identify NCH organizations at risk, and will also hopefully serve as a tangible introduction to linked open data for cultural heritage workers of all kinds.

4. Lessons learned – what insights did your organization gain while implementing the project? These could be insights into such areas as the community served, your own organization, or issues your project sought to address. Lessons learned are often unexpected.

First, we quickly realized what a strong community of cooperation Georgia has for responding to disasters. We were able to tap into an existing network of passionate preservation and conservation professionals across the state, so we were met with enthusiasm at the prospect of creating this tool. This may not be the case for other states or regions, but it certainly was a pleasant surprise that made getting the initial datasets much easier.

One unfortunate component of this project that was unanticipated was the emotional toll it took on us. For context, this project was completed at the same time as international protests were taking place against systemic and police violence committed against Black Americans. Georgia's history includes the history of slavery in the South, and many of the records that we saved, enriched, and updated reflect this reality. There are organizations who celebrate those that defended the practice of slavery, and there are also organizations who celebrate those that sought to abolish it. Seeing this same battle play out on American streets while we worked on these records was uncomfortable, to say the least. Yet, no matter how we may personally feel about parts of Georgia's history, we endeavored to record it all fairly and equally. No record should be erased because we are uncomfortable. Our discomfort should push us into discourse about what each record means to the state of Georgia then and now.



A more pleasant surprise in working on the project was how Wikidata was being used by others to store and discover additional information about the NCH organizations. During the project, a Wikidata user uploaded visitor counts for all Georgia Public Library System libraries from 2015-2017. Because this data was well structured, sourced, and stored in Wikidata, we were able to graph the visitor counts for the entire state's public library system in a matter of minutes [see graphic to left] (https://w.wiki/QDW). Since Wikidata is an open system, additional information can be added to these records, such as budget changes, collection statistics, branch openings and closings, and other facts, all stored as linked open data. Likewise, the data that we entered for disaster preparedness and recovery is now freely available to be repurposed for tourism, planning infrastructure, performing historical analysis, or other topics we have yet to imagine.

5. What's next – where do you go from here? How do you plan to sustain and/or expand upon your accomplishments both for your organization and the community? Is there a way that LYRASIS can further support your efforts in continuing the work, or play a role in taking this further in some way?

We recognized early on that this project would need a state-level institutional home, therefore we created a sustainability plan with our partners for inclusion into the project. GALILEO, Georgia's virtual library, has graciously provided website hosting for the GaNCH website. The Georgia Public Library Service's Archival Services and Digital Initiatives Office and GALILEO have both volunteered to provide ongoing staff support for dataset maintenance via annual updates.

To help ensure long-term sustainability of the project, we have built a semi-automated emailer on the back-end of the GaNCH website to send out reminder emails to NCH organizations. Since we gathered organizations' contact email addresses during the enrichment stage of the data workflow, update reminder emails can be sent annually to confirm accuracy of, or obtain updates to organizational information. If inaccurate, the NCH organizations can email back to provide updates. If the email bounces, then staff can check to see if the organization still exists.

LYRASIS can further support this project by funding similar projects in other states. We believe the biggest barrier to success in a project like this is funding for dedicated staff/student/web developer time.

6. How can others find out more?

In order to make this project a model for other states, we have made all project documentation and materials freely available via GitHub: https://github.com/clifflandis/GaNCH.

The GaNCH website is available at: https://ganch.auctr.edu/

7. Are there any examples of PR, presentations, recognition or marketing materials that were produced during the course of the project? If so, please share those.

We've gathered all press releases, flyers and recording of presentations (including several local and national conference presentations detailed below) on the project's GitHub site here: https://github.com/clifflandis/GaNCH/blob/master/docs/press.md.

- **Conference Lightning Talk**
  "GaNCH Project Overview," Society of Georgia Archivists Annual Meeting, Augusta, GA, 10-07-2019
- **Webinar Presentation**
  "Understanding Our Heritage: The Critical Importance of Mapping our Cultural Resources," American Institute for Conservation, *GaNCH case study starts at 34:45*, 02-19-2020
- **Webinar Presentation**
  "Ready for the Worst: Innovative Tools for Disaster Preparedness in the GLAM Community," Heritage Emergency Response Alliance (HERA), 05-06-2020
- **Conference Presentation** (virtual)
  "Using Linked Open Data for Georgia's Natural Cultural and Historical Organizations Disaster Response," LD4 Conference on Linked Data in Libraries, 07-23-2020
- **Conference Presentation** (virtual)
  "Zooming In: Leveraging GIS and Linked Data to Protect Cultural Heritage at Home, Across the US, and Around the Globe," Society of American Archivists Annual Conference, 08-08-2020
- **FORTHCOMING:** 2020 Digital Library Forum, 10-2020
- **FORTHCOMING:** 2020 Society of Georgia Archivists Annual Meeting, 11-2020
- **FORTHCOMING:** LYRASIS Catalyst Project Webinar, TBA