< issue toc next >>

Geospatial Data Management Interview Series: Interview with Stephen Appel

Editor: Sarah Zhang

Welcome to the third interview in this Geospatial Data Management Interview Series! This series aims to provide current, practical information on library projects and practices in managing, stewarding, and curating geospatial data. Any suggestions are welcome; please send them to me at s_zhang@sfu.ca.

Please enjoy the following interview with Stephen Appel, Geospatial Information Librarian at UW-Milwaukee Libraries.

About AGSL GeoDiscovery

American Geographical Society Library (AGSL) GeoDiscovery is the University of Wisconsin-Milwaukee Libraries' instance of GeoBlacklight, a community-developed and federated geoportal. To UWM students and researchers at the American Geographical Society Library, AGSL GeoDiscovery is a tool for finding GIS data, digitized maps, web apps, and geospatial web services from multiple sources. The Geoportal provides access to AGSL geospatial data collections, maps and data from OpenGeoMetadata, strategically harvested Wisconsin open data resources, and licensed datasets requiring institutional access. The UWM Libraries and the AGSL started with strategic goals to improve access to geospatial resources and identified GeoBlacklight as a core component of our solution because of its community and fit for our unique needs as a geography research library at a public, urban research university.

Q: Could you give us a bird's-eye view of how

AGSL GeoDiscovery is structured? I understand that the discovery layer uses GeoBlacklight, but do you rely on a file server or a repository to store the data layers? How do you generate web services for those data layers?

Our GeoBlacklight backend is quite straightforward: an Apache server with zipped archives of datasets. We have just under 1,000 datasets from our own collection indexed in GeoDiscovery. There are still some low-use, difficult-to-document datasets that have yet to be processed.

We don’t generate web services for our data, not yet anyway. The front-end functionality is there, as it’s built into GeoBlacklight. Data we harvest from other institutions via OpenGeoMetadata or datasets we scrape from DCAT sites with associated services will appear in the preview. Our data includes simple bounding boxes. In future planning, we’ve considered a GeoServer instance or perhaps leveraging our ArcGIS server or ArcGIS Online for hosting web services, but we haven’t yet begun moving in this direction.

We had investigated using Samvera as a backend to manage our data, but we decided it was overkill for our use case. We were hoping that it would provide some tools to help with digital preservation tasks.

Our restricted datasets, available only to UW-Milwaukee or other Universities of Wisconsin users are managed by placing those download links behind our university system’s single sign-on. This allowed us to avoid adding authentication to our GeoBlacklight instance.

Q: How big is the team responsible for the ongoing development, maintenance, and data curation of AGSL GeoDiscovery?

It’s small. I’m the only person doing the ongoing development and calling it “development” is being generous. I’m trying to add new plugins like Blacklight::Allmaps (https://github.com/bplmaps/blacklight-allmaps) for displaying

georeferenced maps. I keep up with my Dependabot updates for dependencies.

Managing the project in GitHub has quite a few advantages, including Dependabot, CI testing, tracking and documenting changes, and the ability to easily share our code with people when we need help.

I have support from a few other UWM Libraries staff—notably our Digital Collections and Initiatives department. They have a systems engineer who splits his time with other units on campus but administers our Linux development and production environments for the library. He worked closely with front end development during the original deployment. He works with campus IT for security, storage, and server resources. I also have the support of a metadata specialist and an application developer who were active in the initial development and planning. I've also had some help from interns and fieldworkers from our library school who contribute to metadata.

I emphasize that a big breakthrough was hiring Eric Larson to do the front-end development and deployment. This required a budget, so, credit also goes to Marcy Bidney, our former curator at the AGSL, for securing funding to make that happen.

The Geo4Lib Community feels like part of the team, too. My number one piece of advice for the GeoBlacklight-curious is to start attending the monthly Geo4Lib community calls. The community is one of the major factors in our decision to use GeoBlacklight. All the way back in the planning stages, we were leaning heavily on community members who were generous with their time and expertise.

Q: I noticed that there is a large AGSL digital map collection as part of your library's Digital Collections. Are these digitized maps also available through AGSL GeoDiscovery? If so, how are these two systems connected?

Getting our digital map collection indexed into AGSL GeoDiscovery is on my list of future enhancements. The sticking point is generating bounding boxes to enable the spatial indexing functionality of GeoBlacklight. It's an interoperability issue; bounding boxes are stored in the MARC catalog record for many, but not all, maps. Our digital map collections are hosted on CONTENTdm, but that metadata usually

does not include the bounding box. Generating an OGM Aardvark metadata record for ingest into GeoDiscovery requires joining these two disparate metadata resources, which is easier said than done given the complicated nature of map description (map series, maps on multiple sheets, compound objects in digital collections, etc.)

All that said, we are working on it. It's a little frustrating to use our own geoportal and seeing maps from other institutions but not our own. Maybe when WAML members are reading this interview, I will have made more progress!

Q: In addition to hosting AGSL's collections, is collecting, archiving, and providing access to data from Wisconsin state and local agencies (including those that already have open data portals) also a motivation for developing AGSL GeoDiscovery?

Yes, absolutely. We used to call ourselves the Spatial Data Clearinghouse for the university, and we want GeoDiscovery to be a solid first stop for students and researchers looking for GIS data. In Wisconsin, counties are important producers and providers of public geospatial data (This was what I studied for my geography MS Thesis, *Public Geospatial Data in Wisconsin: Information Access, Data Sharing, and the University*). Our colleagues at UW-Madison have been working for years now on collecting core datasets from all 72 Wisconsin counties and making it available on their own geoportal, Geodata@Wisconsin. Before we had a geoportal of our own, Geodata@Wisconsin was primarily where I directed our users to find Wisconsin open GIS data.

We ultimately decided that we wanted to host data beyond Wisconsin that better reflects AGSL's collecting scope and UWM researcher needs, including datasets that are restricted or licensed. Even before GeoDiscovery, we made an effort to collect and archive important local datasets from state agencies, municipalities, and quasi-government organizations like regional planning commissions and utilities. But automating that process has ensured that our users are finding up-to-date resources directly from the source. After my project to incorporate our digital map collection, my next target is automated archiving and snapshot capture of

important services.

I started with some python code from UW-Madison to scrape from DCAT-enabled sites, like ArcGIS Hub sites. However, the versatility of the DCAT schema means that not every resource is something we want to link to, and that requires a lot of manual intervention to sort through irrelevant resources and supplement metadata.

One practical tip: If you want to explore what a DCAT repository looks like behind the scenes, open up the DCAT Feed link in OpenRefine and take a look at the JSON data. This helped me to refine the python scripts and look for common issues like missing or template field values and identify systematic issues with geography (like missing, incorrect, out-of-order bounding box coordinates).

Q: Can you share a few major challenges in maintaining AGSL GeoDiscovery?

After months in production, GeoDiscovery started to crash and we were starting to get downtime alerts on a weekly basis. Then it was daily and even hourly. I had been reading about others who were experiencing an influx of web crawler traffic.

I’m no expert, but my understanding is that web crawling is an important part of the internet. Google, for example, is crawling web pages to help give searchers relevant, up-to-date search results. But these crawlers respect limits set up in robots.txt files and crawl in such a way as not to overtax the resources. So called AI crawler bots have been less well-behaved and tend to make repeated, complex searches, choosing multiple facets, and causing Solr to crash under restrained memory resources.

I’m hoping that the GeoBlacklight, Blacklight, and other community-developed projects continue to collaborate on ways to deal with this issue. We had an unconference session about solutions at last year’s Geo4LibCamp and this resulted in a few institutions, including ours, using Cloudflare’s free Turnstile widget. If you’ve been on the internet lately, you’ve seen Turnstile— “Verify you’re human!” Campus IP ranges and some other known ranges are safe listed to avoid friction for our target users. We also do temporary, targeted geographic blocking where the bots are connecting from (China, Brazil, and Singapore in our experience). The

combination of these responses has dramatically reduced downtime, which we monitor with UptimeRobot.

Q:
AGSL GeoDiscovery uses OpenGeoMetadata Aard vark as its metadata framework. Is converting datasets that used other metadata frameworks a significant task?

The AGSL was interested in the GeoBlacklight (pre-OGM Aardvark) metadata schema for longer than I've been in my position. We undertook a massive metadata authoring project, always with the intention of making metadata that could be translated to GeoBlacklight-friendly metadata using FGDC and ISO standards. ISO is our canonical, descriptive metadata record for most of our datasets and the OGM Aardvark JSON files are what we expose on OpenGeoMetadata so others can harvest.

The actual conversion isn't all that arduous. This is where having a metadata specialist on the team has been crucial. He set up a workflow in Oxygen XML Editor to process our XML metadata into OGM Aardvark JSON files. We've done some experimental projects where we're harvesting datasets, writing Aardvark directly in spreadsheets, and using python scripts adopted from other community members to generate the JSON files. I imagine our map metadata workflow being similar once we finalize it since the canonical map metadata is managed by our map cataloger.

One of the biggest challenges related to metadata is ensuring that the metadata we harvest from others via OpenGeoMetadata plays nice with our geoportal. For example, other institutions may implement fields in a different way or use custom fields not found in the schema. Some of my scripts have complex loops that incorporate customizations and overrides for specific institutions.

Q: What are the most viewed or downloaded data or maps from AGSL GeoDiscovery?

We don't collect this information actively, and I'm not sure we have the volume for it to be super useful yet. We do have Google Analytics wired up but we don't

analyze it regularly. I can say that we've had a lot more external users from the public than I anticipated. Oftentimes, their questions are about data I harvested from other institutions and I'm able to direct them to a data source they can access.

Q: Can you share a bit about your future plans or anything that particularly excites you?

We're keeping an eye on the development work happening over at the Big Ten Geoportal, specifically the OpenGeoMetadata API. I agree with many in the community that the most important product of the whole GeoBlacklight project has been the schema and I'm happy to see the community moving towards a metadata API structure. This will allow more flexibility in what people use for their front end, whether that's GeoBlacklight or something else.

I'm also super excited to see the rebranding of the GeoBlacklight Community to the Geo4Lib community. There's been lots of interesting projects and topics like Allmaps, Cloud-optimized GeoTiffs, PM Tiles, collaborative georeferencing, and OpenIndexMaps that are only tangential to GeoBlacklight but have lots of interest and expertise overlap.

Karen Majewicz at the University of Minnesota deserves so much credit for the community building she's done with that group. I would encourage anyone reading this, even if GeoBlacklight isn't on your radar at all, to check out one of the monthly Geo4Lib community meetings (https://www.geo4lib.org/events). It's a great mix of developers and librarians who are all interested in geospatial libraries and technology.

If any WAML members want to chat about my experience or have questions, I'm happy to pay it forward and talk over the challenges we faced and how we've been working through them. You can see ASGL GeoDiscovery at https://geodiscovery.uwm.edu. You can see our internal documentation at https://uwm-libraries.github.io/GeoDiscovery-Documentation/.

< issue toc next >>