Web Services Report

Introduction

Initially called the "ArcGIS Enterprise Organization Working Group", our focus was investigating how the BTAA Geoportal project might utilize ArcGIS Online as a platform to create web services for displaying geospatial data previews. However, after the survey of Associate University Librarians indicated interest in exploring geospatial web services across the participating BTAA Geoportal institutions, it became clear that we needed to expand the scope of our work to include a broader discussion of web services, technologies and workflows.

Understanding the basics of what web services are and how they are created brings to light potential pathways forward and a clearer understanding of the work involved. Because of varying levels of access to the primary platforms used in creating web services, we approached our technology research in two ways. We reviewed ArcGIS Online workflows using a hands-on approach, accessing the Big Ten Academic Alliance ESRI Organizational account directly. However, in researching the open-source platform GeoServer, we relied on informative interviews and experiences from experts at Princeton University. The following report provides information about what web services are, how they might benefit geoportal users, what technologies might be used to create them, how collection development and metadata are impacted, and the advantages or disadvantages to different approaches of implementation.

• What is a geospatial web service?

A geospatial web service is a GIS resource (e.g., data, map) that allows users to preview, visualize, and repurpose data in web maps and applications without having to download or host a new copy. Geospatial web services depend on a server-side / client-side relationship. That server may be on premises or may utilize cloud-based infrastructure, and those clients can access, combine, and (in some cases) contribute information back to web services from different sources without the need for local hosting on their own infrastructure.

Different types of geospatial web services exist, with two of the most common and relevant for our purposes being Web Feature Services (WFS) and Web Map Tile Services (WMTS).

1. A web feature service is a dynamic, read-only service useful for directly exposing data for display and querying in a web map or web mapping application. Example: Web Feature Service.

2. A web map tile service is a dynamic, read-only service that supports fast visualization from a collection of predrawn images (tiles) but does not allow direct access to data about individual features. Example: Web Map/Map Tile Service.

The role of web GIS in the geospatial industry continues to expand and evolve, and it is in this context that geospatial web services have become an important mechanism for distributing geospatial information. By linking to a web service hosted by an authoritative data provider, users are able to display the most accurate and up-to-date information available in their web maps and applications. This decreases duplicate versions that may become out of sync if corrections or updates are made by the data provider. Geospatial web services can also pose a challenge though for users seeking to directly access the data on which those web services have been built, as a user's ability to download those data may be limited by the settings configured by the owner of that web service or, when the configurations allow for data querying/download, may require additional skills to carry out the necessary steps in a desktop GIS.

Benefits of web services

Within the context of the BTAA geoportal project, providing web services would have a number of potential benefits:

1. ***Ability to preview data before downloading.*** Web services allow users to view the spatial extent and attributes of a dataset without needing to open it in a geospatial software. Example: Bike Lanes: Detroit Michigan
2. ***More consistent display and user experience for geospatial data.*** A large number of the geospatial datasets available through the geoportal already have associated web services. Hosting web services for data that do not currently have them would support a more consistent user experience for records from across institutions and data sources.
3. ***Incentive for researchers to use the geoportal.*** Creating geospatial web services from data hosted in institutional repositories could benefit researchers who do not otherwise have a stable place to visualize their data. This could also possibly benefit collection development by increasing the number of unique datasets available through our geoportal.

Roughly half of our current geospatial datasets do not have an associated web service, although the percentage varies between institution:

Geospatial Data Records in the BTAA Geoportal (Oct 2019)

Institution	With web services	Without web services	Percent w/o web services
Wisconsin	41	2,902	99%

Michigan	11	106	91%
Penn State	436	1,515	78%
Minnesota	347	710	67%
Chicago	168	285	63%
Illinois	69	120	63%
Maryland	2,111	824	28%
Iowa	854	99	10%
Purdue	273	18	6%
Ohio State	1,107	27	2%
Indiana	159	0	0%
Michigan State	450	0	0%
Total	**6,026**	**6,606**	**52%**

Overview of web service technologies

We investigated two options for hosting web services to compare both open source and proprietary solutions - GeoServer and ArcGIS Online respectively.

GeoServer (http://geoserver.org/)

GeoServer is an open-source Java application server that supports the development, hosting, and sharing of dynamic map and data online services. It is an Open Geospatial Consortium (OGC) compliant implementation of a number of open standards such as Web Feature Service (WFS), Web Map Service (WMS), and Web Coverage Service (WCS). Additional formats and publication options are available including Web Map Tile Service (WMTS) and extensions for Catalogue Service (CSW) and Web Processing Service (WPS).

Resources and workflows

As an open source platform, there is abundant documentation to provide guidance in implementation, development, and customization of GeoServer. One major benefit of this option would be the ability to test new releases and features before deciding to implement them locally. While it does require dedicated maintenance, the web API is quite simple and there is an intuitive graphical user interface (GUI) to store and manage data and create

services. There is typically one dedicated "workspace" where map layers are created -- both raster and vector data can be stored and hosted as services. There are very few (if any) file size limitations that would hinder certain datasets from being shared as a service (the limits that do exist are aimed at maintaining reliable draw performance/speed). Online documentation and an active user group within the GeoBlacklight community are available to offer guidance on automation and scripting for more efficient workflows. There is support for metadata within GeoServer -- simple records can be added in the GeoServer workspace, and more complex metadata (if available) can be uploaded separately. Unlike ArcGIS Online, Geoserver is not itself a discovery platform, so it would not be necessary to transform and maintain metadata records to align with an additional schema. Existing BTAA GeoBlacklight metadata could be used to populate descriptions of web services.

Implementation - distributed vs. centralized

Although the technical requirements and maintenance of GeoServer are not especially demanding, there must be dedicated IT support and server authorization permissions. Data files are stored in a database (i.e. PostGreSQL) with a client interface for uploading content. There are some advantages and challenges with both distributed and centralized implementation of GeoServer.

Distributed

The notion that each participating BTAA institution would stand-up separate instances of GeoServer may not be practical, but it is not impossible. Unlike ArcGIS, GeoServer is not an already widely utilized tool at BTAA institutions. It would have to be implemented as an addition to each University Library's existing technical infrastructure. However, a distributed model: (a) allows for local control over storage and management of the geospatial data needed to create web service layers; and (b) reduces negative performance issues by coupling the database environment with GeoServer locally. It might also mean, though, that certain institutions may implement GeoServer and generate services for their locally-hosted data, while others simply do not have the IT infrastructure, expertise, or personnel to do so.

Centralized

It could be possible for a single institution to implement an instance of GeoServer that others could be given permission to access -- but in a centralized model, challenges arise in the storage of data files used in the creation of the services themselves. Geospatial data files must be stored (or archived) for a web service to be generated from them. A single GeoServer instance would struggle to function adequately when accessing data stored locally at institutions across the BTAA. Based on an informative interview with Eliot Jordan at Princeton University (where they are actively using GeoServer and GeoBlacklight) service performance and IT security issues arise in a distributed environment. It is recommended that data files used to generate web services be stored in one location that is directly connected to

GeoServer. This poses potential challenges for BTAA institutions in transferring data from one institution to another for the purpose of web service creation.

ArcGIS Online

(https://www.esri.com/en-us/arcgis/products/arcgis-online/overview)

ArcGIS Online is a proprietary, cloud-based, enterprise web GIS platform. Organizational accounts for teaching and research use of ArcGIS Online are supported through Esri educational site licenses, though the number of named users, analysis/storage credits, and account management roles and responsibilities may differ across institutions. Whereas GeoServer is primarily an application for hosting and sharing geospatial web services, ArcGIS Online also serves as a discovery tool and visualization platform for creating maps and analyzing data. ArcGIS Online supports OGC-compliant web feature services (WFS), web map services (WMS) and web map tile services (WMTS), among others.

Resources and workflows

There is abundant documentation available for publishing hosted layers and web services to ArcGIS Online. Organizational accounts come with access to Esri tech support services as well as to Esri training catalog resources. There are several different workflows for creating and publishing web services depending on the desired functionality and the type of service being created. For hosted feature services, one key consideration is a workflow for updating the maxRecordCount property, which defines the number of features that can be accessed when querying a web feature service or when downloading to create a local copy. For hosted tile layers, the recommended publication workflows are different depending on the size of the data set (less than or more than 1GB). While these are not file size restrictions per se, they are important to keep in mind for ensuring appropriate access to and efficiency of the hosted datasets in ArcGIS Online. The metadata workflows in ArcGIS Online are a bit more limited and more complicated than those for GeoServer. Currently, only the ArcGIS metadata format is supported for importing into ArcGIS Online. Furthermore, the specific publication workflow utilized for adding hosted layers to an ArcGIS Online account will influence the way that metadata is included with those items (see FAQ here for more detailed information about metadata publishing workflows).

Implementation - distributed vs. centralized

While most contributing institutions have access to an institutional ArcGIS organization, task force members have different roles, responsibilities, and permissions in relation to their local instance. If a task force member moved to a different job, they may not be able to continue to host the web services they had uploaded to their institution's ArcGIS organization. Changing ownership of content, especially between organizations, is challenging. Because of this, we recommend web services generated specifically for the geoportal be uploaded to a centrally administered BTAA ArcGIS organizational account.

Creating and distributing geospatial web services with ArcGIS Online would utilize Esri's cloud infrastructure rather than any local IT solution, so the distributed vs. centralized technology considerations outlined above for GeoServer do not necessarily apply.

Task force members would be able to upload data and create services within the BTAA ArcGIS organizational account from dispersed locations. This could work well for research data held in institutional repositories, in cases where substantial curation is needed and in person communication with researchers preferable. In other words, storage would be centrally managed, but the labor of creating certain types of web services would be distributed across task force members. For some types of data, it may make sense to use centralized workflows. Currently much of the public/open government data is accessioned by central project staff through automated processes. Trying to coordinate geoportal records generated through these processes with web services created by distributed task force members might prove to be complicated.

As the individual institutions participating in the BTAA geoportal project also have our own ArcGIS organizational accounts to support research and teaching, there is a separate question in regards to scope, i.e., which datasets/services might be hosted centrally in the BTAA organizational account vs. which may be hosted locally at our institutions. One way of approaching this could be based on if the geoportal task force member is directly responsible for creating the web services or if researchers at our respective institutions are creating and sharing the web services within our local ArcGIS organizational accounts. In the former case, the BTAA organizational account likely makes the most sense, especially as individual task force members cycle on or off of the project. In the latter case, the datasets/services may be hosted in local ArcGIS Online organizations, but records could still be created in the geoportal to facilitate discovery and access for these data.

Collection development considerations

Considerations for collection development as it applies to web service development would likely be undertaken by individual institutions (task force members) with guidance from the BTAA Collection Development Committee. Each institution would play a significant role in evaluating existing and potential content for appropriateness as a web service.

Initial collections to consider may include:

- Geospatial data housed within academic institutional repositories that do not have web services
- Public or open geospatial data that does not have a web service
- Public or open geospatial data that is no longer being hosted by the original data creator or has a web service available from an unstable source

There are differences in the approaches to creating web services for research data (i.e. housed within an institutional repository) vs. public/open data published by state or local governments. Research data is information already stored, managed, and preserved by a University -- making them a (potentially) more stable source of content for web service generation. Public/Open data from state and local governments would need to be acquired, curated, and stored locally at an institution to enable creation of web services. Some specific challenges include:

Research data (institutional repository):

Curation

Geospatial data within institutional repositories are available in a variety of formats and can be stored within complex research project structures. Decisions must be made item by item about how best to represent this geospatial data as a web service. It may not make sense to create web services for all spatial research data (such as water quality sampling locations for a specific study.)
Example: Characterization of streams and rivers in the Minnesota River Basin Critical Observatory: water chemistry and biological field collections, 2013-2016

Multiple data layers for a single record

Data repositories often store all data layers related to a project together. These projects are currently represented in the geoportal with a single record. It would be difficult to link multiple web services to a single geoportal record, requiring a more complex representation of these resources.
Example: Access Across America: Transit 2015 Data

Licenses

Licenses assigned to research datasets may impact whether they can be re-distributed as a web service. Datasets assigned public domain, CC0, or CC-By licenses may be appropriate as an initial focus, whereas CC-SA or CC-ND could be more complicated. If the web services are being hosted from a centralized location, there are also could be questions around sharing research data between institutions.
Example: White-tailed deer density estimates across the eastern United States, 2008(Attribution-NonCommercial-ShareAlike 3.0 United States)

Special capacities (e.g. time-enabled layers)

Some data layers would be most useful (and true to their original purpose) if time-enabled, but it would take more work through more complicated workflows to create these web services.
Example: Reconstruction of North American Drainage Basins and River Discharge Since the Last Glacial Maximum

Communicating with researchers/students/original data creators
Creating web services will require cartographic decision-making around the best way to display each dataset as a map or image. Communication with the original data creator may be necessary to understand the content and how best to represent it..

Public/Open Data:

Syncing web services with data records hosted by original data creators
Web services created by the BTAA Geoportal project would be representations of point-in-time copies of datasets from original data creators. This means the ability to sync a service with the version of the dataset hosted by the original data creator will be difficult. As original data creators update, edit, or remove content from access online - there will be challenges in tracking what content we have generated web services for and whether or not they also need updating, editing, or deletion. The BTAA project essentially becomes the steward of a particular version of a public/open dataset when creating a web service for it.

Curation
Creating web services for public/open data produced by state and local governments means a local copy of the data must be stored and hosted by the participating BTAA institution(s). Content from original data creators would need to be reviewed item by item to determine the need for a web service.

Licenses
There are some cases where licenses are not clearly assigned by the original data producer. Navigating access and use constraints for data with ambiguous licenses means direct communication with data creators may be required. Our current metadata schema does not include liability statements or licenses. If we were to become stewards of this data, we may need to expand the metadata elements we keep.

Communicating with original data creators
There may be a need for a different model of communication with data creators if the BTAA Geoportal project plans to copy and display the original data as a web service. The current model consists of simply adding an additional discovery mechanism to the information data producers are hosting themselves. Conversations on the best way to represent the data (cartographically) as a web service may also be necessary. Formal channels of communication regarding data stewardship and representation may need to be established in a way that has not yet been fully explored by the BTAA project.

Metadata and workload considerations

In order to add web services to our current geoportal workflow, more attention would need to be given to the metadata for those resources. In addition to the Geoblacklight metadata currently being created for display in the Geoportal, we would need to locate and process copies of more comprehensive metadata when available. Currently we rely on the original data source to provide supplemental information needed to make sense of data layers. But if we are keeping our own copy of data that will persist independently of the data provider, we will need to be more attentive to this information. Comprehensive metadata is much less standardized and complicated than the more limited Geoblacklight schema, meaning that these steps will likely require a substantial amount of manual labor. We will also need to track actions related to the stewardship of the individual data files. We also might need to create a second set of discovery metadata for each web service if using the ArcGIS Online platform to host. It is unclear who would be responsible for this work since it would be challenging to standardize these workflows steps between our distributed task force but time-intensive if performed by a centralized staff. Below is a list of potential additional steps beyond our current process that may be needed to offer geospatial web services:

- Communicating with researchers or data providers to clarify ambiguous licenses and to select appropriate cartographic representations for the data
- Downloading a snapshot of data to serve as the local copy supporting the web service
- Creating additional metadata including administrative information tracking the download / processing of the web services data and relationship to the Geoblacklight metadata and (if hosting in ArcGIS) ArcGIS Item Description metadata
- Creating web services either by adding to a PostGreSQL database or uploading to ArcGIS Online
- Adding new web service links to the related Geoblacklight metadata
- Tracking modifications to the original dataset and making decisions about whether to reaccession to stay in sync or to continue linking to the snapshot

We estimate that these additional steps could add 30 minutes to the processing time for an average item.

Conclusions / Recommendations

Creating web services would provide a more consistent and satisfying user experience for interacting with geospatial data in the Geoportal. In particular, it allows users to preview data even if they do not have access to geospatial software and before downloading files to their local device. It may also encourage meaningful connections with faculty and students

interested in ways to visualize and promote their research data. Expanding the scope of the project in this way would require, however, a substantial increase in work for task force members and project staff. Beyond the current process of creating discovery metadata for each data layer, the project would need to download a copy of each dataset, gather more detailed metadata, generate and link geoportal records to web services, and track modifications to the original dataset. We estimate this could add up to 30 minutes to the processing time for the average item.

One of the more important decisions to be made will be whether to take a more centralized or decentralized approach to creating web services. If hosting content through Geoserver, a more centralized approach is recommended. This is because the services would need to be supported by a local database and it would be inefficient to set up servers at each partner institution. If hosting through ArcGIS Online, a more decentralized workflow may be feasible. This is because data and services would be stored in the cloud and any task force member would be able to upload data to the shared BTAA account. More research needs to be done on the specifics of hosting using GeoServer before a recommendation can be made on the best technology option. While GeoServer would take a greater infrastructure investment to set up and maintain, it would give us the greatest control over web service resources and the way they interact with the Geoportal. ArcGIS Online would be comparatively easy to set up and administer, but we would have less control over performance and storage. Also, since ArcGIS Online is itself a discovery platform, we would have to maintain metadata for both our Geoportal and their interface.

Creating web services would introduce a new set of challenges regarding how to maintain and represent data in the Geoportal. For research data, there would need to be significant curation and item-by-item decision-making about how to best represent each project. For public data, the most formidable problem will be how to keep web services accurate and non-duplicative as original data creators update, add, and remove content. In both cases, creating web services may require a kind of engagement and conversation with data producers that we have previously skipped.