Collaborative Librarianship

Volume 9 | Issue 3 Article 7

11-21-2017

A Spatial Collaboration: Building a Multi-Institution Geospatial Data Discovery Portal

Mara Blake
Johns Hopkins University, marablake@jhu.edu

Karen Majewicz
University of Minnesota Librarianes, majew030@umn.edu

Ryan Mattke
University of Minnesota Libraries, matt0089@umn.edu

Kathleen W. Weessies
Michigan State University, weessie2@msu.edu

A Spatial Collaboration: Building a Multi-Institution Geospatial Data Discovery Portal

Cover Page Footnote

Acknowledgements The authors would like to thank current and past project Task Force members (Paige Andrew, Linda Ballinger, Kevin Dyke, Cathy Hodge, Melinda Kernik, Nicole Kong, Shirley Li, Jennifer Liss, Jaime Martindale, Kelley O'Neal, Bria Parker, Nathan Piekielek, Theresa Quill, Nicole Scholtz, Rob Shepard, Amanda Tickner, Tim Utter, James Whitacre, and AJ Wortley), the Project Sponsors (Kim Armstrong, John Butler, Cliff Haka, James Hilton, Wendy Lougee, and Claire Stewart), the Web Development Team at the University of Minnesota Libraries (Michael Berkowski, Paul Bramscher, Cody Hanson, David Naughton, and especially Eric Larson), and Len Kne for his input on the Project Proposal. A very special thank you goes to Mark Sandler, without whose support this project would not have happened at all.

A Spatial Collaboration: Building a Multi-Institution Geospatial Data Discovery Portal

Mara Blake (marablake@jhu.edu)
Data Services Manager, Sheridan Libraries, Johns Hopkins University

Karen Majewicz (majew030@umn.edu)
Geospatial Project Metadata Coordinator, Wilson Library, University of Minnesota Libraries

Ryan Mattke (matt0089@umn.edu)
Map & Geospatial Information Librarian, Wilson Library, University of Minnesota Libraries

Kathleen W. Weessies (weessie2@msu.edu)
Geosciences Librarian / Head, Map Library, Michigan State University Libraries

Abstract

As academic education and research increasingly take advantage of geospatial data and methodologies, we see a corresponding exponential growth in the number of available geospatial resources in the form of GIS datasets and scanned historical maps. However, users can experience difficulty finding these resources due to the unconnected multitude of platforms and clearinghouses that host them. Additionally, the resources are not always well described with web semantic metadata that facilitates discovery. In response to this challenge, The Big Ten Academic Alliance Geospatial Data Project began in 2015 to provide discoverability, facilitate access, and connect scholars to geospatial resources. Our project leverages a multi-institutional collaboration and open source technologies to improve discovery for users of geospatial data and scanned maps. We outline collaborative workflows and strategies for a successful multi-institution collaboration.

Keywords: geoportals, consortia, collaboration, geospatial, maps, discovery, metadata

Introduction

As academic education and research increasingly takes advantage of geospatial data and methodologies, we see a corresponding exponential growth in the number of available geospatial resources in the form of GIS datasets and scanned historical maps. However, users can experience difficulty finding these resources due to the unconnected multitude of platforms and clearinghouses that host them. Additionally, the resources are not always well described with web semantic metadata that facilitates discovery.

In response to this challenge, the Big Ten Academic Alliance (BTAA) Geospatial Data Project began in 2015 to provide discoverability, facilitate access, and connect scholars to geospatial resources. Our project aims for the following three goals: 1) A public collection of harmonized, platform-agnostic geospatial metadata; 2) A shared geoportal for institutions across the Big Ten; and 3) Development of workflows and use of tools. The public face of our project, the BTAA Geoportal, offers a single, aggregated interface for users to discover geospatial data and scanned maps from a variety of sources.[1] Our project leverages a consortial collaboration and

open source technologies to improve discovery for users of geospatial resources. (See Figure 1.)

Literature Review

The dispersed nature and lack of standard description methods for geospatial data make it difficult for users to effectively and efficiently discover the resources they need. Data covering the same area may be available from multiple providers on multiple websites that do not reference each other. Data may exist for a topic and geographic area, but require payment or direct interaction with a provider for access. Data may exist, but not be available for public use. Data may not exist at all. Tools like ArcGIS Online allow users to see data that may or may not have accompanying metadata or provenance information. In this landscape, users of geospatial data often find that learning whether or not data exists, and then acquiring access to that data, can be a difficult and frustrating process. For the purposes of this article, we will focus on the landscape in the United States and the role that academic libraries play. We do want to note that Europe's landscape for geospatial data looks very different, where European Union (EU) guidelines led to early work on European Spatial Data Infrastructure (ESDI) and an EU Geoportal.[2]

Libraries in the U.S. have attempted to address challenges in discovering geospatial information. GeoDex was a notable early search interface for geographic collections, invented by Chris Baruth at the University of Wisconsin-Milwaukee in 1988. The university used it in-house and distributed it to customers as software and instruction manuals. Since there was no practical way to crosswalk the information already gathered in catalog record fields, staff members generally typed the bounding coordinates of each individual map sheet (and all the other bibliographic information) by hand. Several libraries pursued the goal of a searchable geographic index only to have it fall by the wayside over time. The American Geographical Society Library, however, quietly carried on with entering bounding coordinates of over 400,000 map sheets creating a body of metadata which can, after 25 years, be utilized for its intended purpose.[3]

Kollen et al. reported on the findings of the Spatial Data Subcommittee of the American Library Association (ALA) Map and Geospatial Information Round Table (MAGIRT) Geographic Technologies Committee, which investigated the response of academic libraries to this landscape.[4] The Subcommittee interviewed 11 institutions, asking about their available geospatial data, discovery tools and technology, staffing, and maintenance issues. The authors found increased support for geospatial data discovery from earlier studies, but reported a great diversity of offerings from the academic libraries interviewed. They recommended that institutions customize their services to their local needs.[5]

To help address this challenge of discovery and access of geospatial data, academic libraries and institutions developed geoportals which serve as aggregators of numerous siloed resources. A geoportal serves as a single, aggregated discovery system for geospatial data. A collaboration led by Tufts University, with Harvard University and MIT, developed and launched OpenGeoportal (OGP)[6] in 2012, the first large-scale open-source geoportal.[7] Florance et al. describe the origins and structure of the OGP Federation.[8] A 2013 Summit funded by an Alfred P. Sloan grant brought together many contributors of OGP and allowed the federation to address many issues, including the development of governance models. The OGP Steering Group governs OGP Federation, operating with a "meritocracy" where those who contribute more have more say in the direction of the project. The Federation utilizes

working groups, notably the Developer Working Group and the Metadata Working Group, to accomplish task-oriented work related to the project.[9]

GeoBlacklight, another open source geoportal option led by Stanford University, went live in 2014.[10] In addition to Stanford, MIT, New York University, and Princeton University all contribute to GeoBlacklight. Stanford runs their own implementation of GeoBlacklight called Earthworks; the GeoBlacklight website shows the many other implementations of the technology.[11] The members of the GeoBlacklight community use a Google group and Slack channel for communication and announcements. Periodically, Geoblacklight developers will schedule "sprints," or condensed time of intense code development.[12] Hardy and Durante introduced the metadata schema that powers the discovery capabilities of GeoBlacklight. They describe the use of metadata schema that is pared down from robust Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC) or International Standards Organization 19139 (ISO) records, retaining just the elements most useful for discovery for search in the geoportal. The full metadata records remain available for users.[13]

The OpenGeoPortal and GeoBlacklight projects resulted in the production of tens of thousands of geospatial metadata records that participating institutions created or gathered. Members of the projects needed a platform for sharing the records in order to make them available for ingest by different geoportals. This led to the development of OpenGeoMetadata, a public repository of geospatial metadata files hosted on GitHub. GitHub eliminates the need to stand up customized technology that might be inaccessible to some users and provides version control to track updates. Since the metadata records are accessible as simple, discrete files via Git or GitHub Desktop, metadata aggregators can easily harvest records into geoportals of all kinds.[14]

Wrangling the metadata proves a central challenge in assembling any type of discovery portal. Web portals can provide different types of search functionality; one type is a federated search, or "metasearch," which indexes existing metadata across external databases to return a list of results.[15] Another type searches an internal set of records that have been harvested or aggregated from multiple sources in advance. Libraries increasingly prefer the second type because of its faster response time, and because it allows for remediation and normalization of the records before they are presented to the user.[16] This clean up is especially desirable in instances where the metadata is limited or pulled from non-library sources, such as commercial publishers or government agencies.

As aggregated metadata portals proliferated, the need to systematically gather metadata from multiple sources arose. This spurred the development of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in the late 1990s/early 2000s, which in turn enabled several large scale metadata aggregation projects.[17] The National Science Digital Library (NSDL) began in 2000 and soon implemented OAI-PMH as a core part of its workflow.[18] The CIC Metadata Portal was an initiative in the mid-2000s with nine universities within the Committee on Institutional Cooperation that aggregated nearly half a million records from nine universities using OAI-PMH.[19] The Digital Public Library of America (DPLA) was conceived in 2010, and has aggregated millions of records from repositories all over the country, most facilitated by OAI-PMH.[20]

Although geospatial metadata can use OAI-PMH, other forms of metadata harvesting protocols address more specific needs of data resources, including geospatial data. These include the Catalog Service for the Web (CSW), an open source protocol from the Open Geospatial Consortium.[21] Many public agencies build their geoportals with the open source Comprehensive

Knowledge Archive Network (CKAN) application, which boasts API plugins that can expose the metadata for harvesting.[22] A rapidly growing interoperable standard is the Data Catalog Vocabulary (DCAT) format, which allows metadata sharing between applications.[23] The proprietary Socrata and ArcGIS Open Data Portal applications utilize DCAT.

Original metadata must meet a sufficient threshold level of consistency and completeness for discovery portals in order to make harvesting effective. Unfortunately, web portal projects frequently deal with inconsistent and incomplete metadata and need to perform significantly more manual enhancements and cleanup of the metadata than they originally anticipated. Discovery portal project teams often begin with a plan for a low-barrier, automated process of aggregating metadata. However, this ideal, streamlined workflow is complicated by the difficulty of harmonizing a myriad of metadata records with varying levels of quality, formatting, and vocabularies. NSDL noted that they were able to automatically ingest a few items, but that "the vast majority of cases required significant manual intervention," due to many different causes, ranging from encoding errors to validation failures.[24] DPLA reported in 2014 that they needed to adjust their metadata profile partially due to multiple challenges of incorporating records from different content providers, and noted that "(i)ngest remains a very hands-on endeavor."[25] They found that spatial terms and geocoding proved especially troublesome, and frequently produced misleading or incorrect results.[26]

Since the aggregation process generally produces such problematic records, institutions have needed to develop remediation strategies. Godfrey and Kenyon from the University of Idaho Library advanced the idea of maintaining a "Geospatial Metadata Manager's Toolbox" comprised of assorted applications and scripts in order to address the many different types of problems that might be encountered.[27] The CIC Metadata Portal project developed a workflow for reprocessing records that consisted of selecting only relevant records for inclusion, removing erroneous characters and empty elements, normalizing the terminology, augmenting the records with additional metadata, transforming them to the format required for the portal interface, and checking performance issues of published records.[28] A thorough remediation process such as this can require extensive time and expensive labor. DPLA has mitigated this obstacle partially by establishing regional service hubs that compile records from local content hubs and ensure that the metadata conforms to DPLA's required profile before submission.

Academic libraries can further contribute expertise to address the challenges with discovery and description of geospatial data. As with projects like DPLA, libraries offer infrastructure and labor to collect metadata records and host discovery systems. Academic libraries also hold expertise in metadata and a deep knowledge of user needs in regards to geospatial data, which they can use to describe data for discovery themselves and provide education to others to author better metadata.

Background

History and Origins of the Project

Our project filled a large gap in information access between geospatial data producers and the academic library world. Unlike the publishing world, where cataloging-in-publication data is placed right onto title pages and publishers eagerly share metadata with union catalogs and indexing services, geospatial data remain closely held and poorly described by the multifold of producers. On the other end, libraries encountered a skill gap; until recently very few academic libraries wrote true geospatial metadata. Map librarians mainly in charge of large paper collections felt unprepared to lead cutting-edge

digital projects, and digital librarians lacked experience with geospatial data issues.

The Director of the Committee on Institutional Cooperation (CIC) Center for Library Initiatives noted this gap in access to geospatial data. The CIC, now called the Big Ten Academic Alliance (BTAA), is the academic consortium of the universities in the Big Ten Conference. The CIC Center for Library Initiatives demonstrated success in negotiating pricing and contract terms for member institutions with library vendors, but also had interest in identifying forward-thinking digital projects. In 2006, the director encouraged CIC map and GIS librarians to form an interest group to explore areas of collaboration around geospatial issues. The CIC map and GIS librarians were generally not acquainted with each other at this point and most were not eager to accept an obligation to a chancy as-yet undefined project. Discussion topics for a few years remained unpressing and pragmatic as the group sought to build trust and a feeling of community. The group considered ideas such as jointly scanning a large body of paper maps and subscribing to map-themed vendor databases.

The developments in cloud-based geospatial metadata indexing described above raised the possibility of collaborating in the area of data discovery. The Open Geoportal project provided a particular inducement, proving that technology and standards were in place to allow a viable, beneficial, and practical project. Discovery of datasets of import to researchers at CIC institutions therefore gained traction as the best area for collaboration.

Three map librarians produced a broad-based white paper describing the needs of geospatial librarians and researchers within the CIC.[29] They noted particular challenges, including increased user demand for geospatial data, limited resources for geospatial services, the high cost of specialized information technology and support for data storage, a limited understanding of geospatial work among generalist colleagues and administrators, and the interdisciplinary and diffuse nature of geospatial data users.[30] The paper also described several collaborative opportunities, including digital collections, services, storage, and access. It noted the potential to develop common scanning strategies for paper maps, to co-invest in technology infrastructure for data storage, dissemination, and archiving, to collaborate on specialized tools, build user communities around centrally supported data resources, and promote the use of geospatial analysis in research and instruction.[31] The team distributed this document along with a proposed blueprint to collaboration, which described people, funding, governance, cooperative infrastructure, partnerships, leveraging existing resources, and administrative support.[32] These documents were distributed ahead of the 2012 CIC Library Conference, which not-coincidentally was themed that year on collaborative strategies for developing geospatial services. The conference also gave nearly all of the CIC map librarians a chance to meet and to discuss in person the white paper and possibilities going forward.

In 2014, a team of three authors drafted a formal proposal, directed at the CIC Library directors, for a collaborative project to develop a geospatial data discovery tool. It articulated a modest trial project in which four institutions would contribute equal funding toward a geoportal. The proposal won a spot on the agenda of a bi-annual director's meeting. The response from the library directors, however, was much more enthusiastic than anticipated, with 11 directors voicing interest in participation. In response, the authors crafted a more ambitious proposal which included a full-time staff member funded entirely by the project, as well as funding for technology infrastructure (web hosting and upkeep), software development (for the open source software platform), travel (for project

members to promote the geoportal), and collaboration (to fund an in-person meeting of the entire group).[33]

Ultimately, nine CIC institutions accepted and funded this revised proposal in February 2015 and the project launched the following July (see Appendix A for a list of participating institutions). The project launched the public BTAA Geoportal in August 2016 with over 2,600 metadata records describing datasets from 21 governmental GIS data portals and university scanned map repositories. As of September 2017, the portal contains over 7,500 metadata records for geospatial data, scanned maps, and web services.

Description of Need

Traditional library discovery tools hamper patrons seeking information based on location in two ways: library catalogs handle geospatial information inadequately and libraries lack experience acquiring and describing geospatial resources.

First, library search indexes and catalogs have historically done a poor job of defining the "where" question. Users can easily search book and journal indexes by title, author, or subject, but face considerably more difficulty in geographically defining an area of interest. When a researcher wishes to find data on a particular location, they may find datasets relating to the area described in a variety of ways. For instance, the area might be described as part of one or more states, as a grouping of counties, by the area served by a regional planning commission, a government agency's own defined service regions, or as being in the vicinity of the various nearby cities. It may be geographically defined by the natural or climatic region, the plants and animals that range in that area, or by the watershed that drains it. A set of geographic coordinates can cut through all these forms of description and clearly mark out the area of interest. Library catalogers envisioned a day when discovery systems could easily use geographic coordinates as a point of entry to searching collections. In 1981, U.S. cataloging practices provided an optional practice for recording map bounding coordinates (in MARC 034 and 255 fields), holding out hope for the possibility that someday libraries would develop a computer-based solution to search catalogs by latitude and longitude.[34]

A second obstacle to users locating data stems from libraries not traditionally acquiring and describing geospatial resources. The richest and most detailed geospatial datasets are created at the local level by units of government, researchers, corporations, and nonprofit organizations. Such data providers generally focus on their own immediate user groups and tend to serve data out in a more casual way than more mainstream data providers. Though book and journal publishers pointedly advertise their datasets in order to facilitate sales, non-commercial geospatial data producers have less motivation and less mandate to make data readily available. Researchers generally located geospatial data by word-of-mouth or by guessing which government, non-profit, corporate, or research entities might have created a dataset they hoped might exist and then contacting the various entities to inquire about available data. A union catalog indexing datasets produced by the full spectrum of data producers would certainly aid the researcher in this process.

One approach to a union catalog is to set a central authority to screen content so that it meets a standard of qualification and quality. The other is a low-bar approach that welcomes large deposits of data files, favoring a large number of easily gathered but potentially lower-quality records. Several projects reach across multiple kinds of data producers to index content. Data.gov brings together datasets from a wide variety of U.S. governmental entities.[35] The RA-

MONA project, now called GIS Inventory, welcomes uploads of metadata from any government agency.[36]

Though many of these projects have been successful in their own way, they have been constrained by variety of data types and lack of maintenance. The academic library world is well positioned to fill this need based on its long tradition of bringing order and standards to the discovery process. The success of such a project rests in striking the right balance between generating custom metadata of high quality and reusing pre-existing metadata which may vary in quality and completeness. Toward that end, our project continually refines its procedures and workflows for gathering and editing metadata.

Collaborative Strategies and Models

Project Structure

Governance

A charter developed at the outset of the collaboration outlined much of the structure and governance of the project, as well as the communication plan and projected timeline.[37] In contrast to the OpenGeoportal model, our project employs a more formal governance structure. The library directors at the three sponsoring institutions and the CLI director serve as the project sponsors and stakeholders of the project. The University of Minnesota hosts the project. The project lead is based at the University of Minnesota and oversees the project, with advice and input from two associate university librarians at the institution. The project hired a full-time project metadata coordinator, who also works out of the University of Minnesota Libraries; the project metadata coordinator coordinates the metadata work across all participating institutions. The project metadata coordinator is the only full-time staff member dedicated to the project. A steering group consisting of three members, one from each of the original university sponsors of the project, directs the project and sets directions. Each participating institution contributes two members to a task force. Those task force members contribute time to the project by: identifying collections of data to include in the Geoportal; acquiring and editing metadata for discovery in the Geoportal; attending monthly task force meetings; and serving on issue-specific sub-groups. These smaller subgroups, made up of members from multiple institutions, review topics and make recommendations to the larger group for general consensus. While the number of records each institution is able to contribute varies by individual circumstances, each participant contributes to the overall direction of the project. In addition to our charter, we developed a document that describes the role of task Force members.[38]

Funding

Each participating institution contributes an equal amount of funding to the budget of the project. The infrastructure of the BTAA provides a helpful mechanism for financial contributions as all institutions have experience transferring funds to the consortium for collaborative licensing and other projects. The funds support the full-time project metadata coordinator, the in-person meetings of the full task force, contract development work, external technology hosting, graduate research assistants, and provide partial support for task force members travelling to conferences to present about the project. Our estimates indicate that the cost per institution to participate in the project is only a small fraction of what it would cost each institution to support a geoportal individually. This financial model is one of the many benefits to pursuing a collaborative geoportal project.

Technology

The project utilizes a range of software applications and scripting languages to transform, edit, and publish the metadata records. Because the

incoming metadata is composed of so many different standards and file formats, we need to employ of a variety of tools to process the records. Desktop applications, including ArcCatalog, MarcEdit, and OpenRefine, are used for transformation and normalization of the records. Two online metadata editors, GeoNetwork and Omeka, are used for collaborative editing of individual records. Python and Ruby scripts are used to batch update, export, and publish records. The geoportal itself runs on GeoBlacklight, which uses Solr for indexing.

The original project proposal called for the use of a cloud-based technology solution for hosting GeoBlacklight. However, within the first six months of the project, it became clear that the project staff did not have the capacity or some of the requisite skill sets to manage a server environment. Thankfully, the University of Minnesota Libraries' Web Development Department stepped up and became a solid project partner. At the beginning of this collaboration, we were just relieved to have someone knowledgeable handling the technical infrastructure, but as the project progressed, members of the web development team became more integrated with our efforts, especially with regards to the areas of interface design and usability. This unexpected contribution to the project resulted in a greatly improved user experience.

Collection Curation and Development

The selection process for records added to the geoportal is informed by a combination of the thematic or administrative calls issued by the Collection Steering Group, by the quality of the metadata as recommended by the Metadata Steering Group, and by the by the ease of which the metadata can be harvested, as determined by the project metadata coordinator.

One of the project's first group activities, before any records were officially submitted, was a survey exploring which collections were most appropriate to include in the geoportal. This included an evaluation of the metadata and an assessment of its priority level. The survey results showed that the most accessible resources with the highest stated priority were GIS datasets from state government agencies and scanned maps held at academic libraries. We decided to tackle the statewide GIS datasets first. Once each institution completed this first round of curation, we turned our attention to scanned maps. Institutions not ready to submit scanned maps instead worked on county and municipal public GIS data.

We have made an effort to make sure that each institution's geographic region is well represented in the geoportal, and that task force members always have a potential collection to work on. This distribution of work has been aided by sharing lessons learned and local practices with each other, such as efficient methods for adding coordinates to scanned map records, stories about how to approach public sector GIS employees or a library IT department, and suggestions for finding lesser known collections of public data.

Metadata Remediation and Workflow

The most critical goal of the metadata workflow is to end up with discovery metadata in a schema that can be loaded into the geoportal. This concise element set is generated by extracting it from a more comprehensive metadata file in a geospatial standard. The geospatial data community typically uses FGDC or ISO, but because FGDC is a legacy standard that is slowly being phased out, we decided that ISO made more sense for long term preservation. However, the ISO schema is so flexible that it can be interpreted in many ways. Without first addressing this variability, any attempts at automated transformations would produce numerous errors.

The first part of the metadata remediation process is structural, in which we ensure that each record is using the same set of elements. Discussions by the task force and within the metadata steering group led to the creation of a standardized crosswalk and templates for default values.

The second metadata remediation task is to normalize the vocabulary. This is particularly important for the most commonly used facet in the geoportal, Place, and all values are aligned to the GeoNames thesaurus. In addition, task force members contribute to a customized synonyms document added to the geoportal's search engine based on their knowledge of common differentiations in spelling for their geographic area, such as "St. Paul" and "Saint Paul," that allows for a better search experience for users.

The project's collaborative metadata workflow utilizes the strengths of all of the participants: task force members are familiar with regional data collections and can spend time making sure metadata records have well-written, quality descriptions, while the project metadata coordinator focuses on batch scripting and troubleshooting errors. (See Figure 2.) Task force members begin the workflow by identifying and collecting metadata records in the form of individual files, harvest links, or a web page of downloadable datasets. The Project Metadata Coordinator then performs various tasks as needed, including harvesting, crosswalking, and batch adding technical and administrative metadata. Once the records have been programmatically normalized through this process, they are uploaded to an online editor. The task force members then log in remotely to edit the records they submitted at the item level by enhancing the descriptive metadata. Finally, the Project Metadata Coordinator publishes the approved records to OpenGeoMetadata and to GeoBlacklight. We repeat this process periodically for each collection to check for new, updated, or deleted records.

Communication

Frequent, short meetings aid the momentum of the project. Our full task force meets remotely on a monthly basis to share general updates and institutional progress reports. The steering groups meet with varying frequencies to make decisions on specific topics. At the technology hub in Minnesota, the project members meet weekly to discuss action items, and the local web developer consults with the project members monthly to plan maintenance and development related specifically to the GeoBlacklight platform. The project lead also meets monthly with the project sponsors to discuss major decisions and future directions. Most of the day to day internal conversations happen online between the project metadata coordinator and individual task force members. All meeting agendas, activity notes, and collection management tasks are documented with G Suite (Google Docs, Sheets, etc.)

For our external stakeholders and the general public, update reports and blog posts are published monthly on the project website or the geoportal blog, and task force members deliver presentations about the project at a variety of venues, including those focused on libraries and others geared towards geospatial data producers. We also collaborate informally with the GeoBlacklight open source developer community to help identify and fix priority issues, and we share our experiences with other institutions interested in creating their own geoportal.

Decisions and Revisions

The project employs an iterative approach where we make adjustments to procedures, policies, workflows, and technology as needed. Over the course of the project, several group decisions have needed to be revisited when we discovered unforeseen problems or when a better option presented itself. We find this adaptive mindset essential to an effective and sustainable collaboration among the participating institu-

tions. In addition to learning from our experiences, we encounter constant flux and evolution in the broader landscape of geospatial data and technology. This approach allows us to adapt to changes and sustain our project.

GeoBlacklight

One example of this approach was the selection and customization of GeoBlacklight. When the project formed, the organizers anticipated using a different geoportal application, OpenGeoPortal. GeoBlacklight came onto the scene shortly after the approval of the project proposal, but before the project officially launched. It offered a better fit, including a more library-centric interface and a more active development community.

Since preparing the metadata records took up the bulk of the task force's early work, the project geoportal initially relied on the default GeoBlacklight settings. The web team at the University of Minnesota customized the home page, but the rest of the pages used default settings. Once the site went public, the project steering group charged an Interface and usability steering group with assessing the user experience. This group conducted a comprehensive usability analysis of the interface, including user testing at three of the participating institutions. The findings of this group led us to make improvements to the interface and enhance the harmonization of metadata for a better user experience. We plan to continue making improvements to the interface and analyze it again in the future. (See Figures 3 and 4.)

Metadata Standards

Our choice of metadata standards provides another example of our flexible approach to revisiting decisions about the project. Regardless of the geoportal technology, the main goal of the BTAA Geospatial Data Project has always been to create a public collection of harmonized, platform-agnostic geospatial metadata. In order to create an interoperable set of records and to facilitate a streamlined workflow, we felt we needed to choose a single recognized geospatial standard for all of the metadata and agreed to use ISO.

However, we revisited this decision when we began to incorporate scanned map records. The academic library community normally catalogs scanned maps with the MARC or Dublin Core metadata standards. In theory, ISO can be used to describe scanned maps, but in practice, it proved to be unwieldy. We did not find a good crosswalk model to translate the MARC or Dublin Core records into the highly nested and codelist-heavy structure of ISO, and several of the required ISO elements, such as topic categories and organizational contacts, proved to be a poor fit for describing maps. The metadata steering group reviewed various metadata records and recommended that we change our plan to continue to use ISO for GIS data, but to use Dublin Core for the scanned maps.

Metadata Editor

The decision to use two different metadata standards also affected our choice of metadata editors. We sought out an online metadata editor with a graphical user interface (GUI). This would enable all task force members to log in remotely to edit their records, and the GUI would lessen the learning curve for working with XML files. We first tried ArcGIS Online, which implemented a metadata editor in 2015. However, we quickly rejected this idea, largely because it does not offer any batch import, export, or updating capabilities. We eventually chose GeoNetwork, an ISO-centric open source application that boasts a GUI and some limited batch editing capabilities out of the box. We extended GeoNetwork's functionality with custom Python scripts that take advantage of the Catalog Service for the Web (CSW) protocol. These scripts give us the ability to batch update many elements with a

spreadsheet, and enable us to export the records into the format needed for the geoportal.

When we decided to work with the scanned map records in the Dublin Core standard, GeoNetwork became a problematic choice. Although GeoNetwork can handle Dublin Core records, we found that harvesting them with the OAI-PMH protocol stripped out several desired elements, and we were faced with a great deal of work to customize it for extended Dublin Core. This challenge led us to a different solution that had been developed at New York University (NYU), namely a geospatial plugin for Omeka. Omeka is well known in the digital humanities field as a web exhibit tool primarily used for digital collections. However, it provides a fairly robust editor for Dublin Core metadata, and features numerous batch editing plugins. NYU created a plugin specifically for creating records for publishing in GeoBlacklight, and included functionality for automatically extracting bounding boxes from GeoNames.

Communications and Project Management

Although the decision to use two different metadata standards and editing tools added a level of complication to the project, we put many of our other adjustments into effect in order to simplify our structure. At the outset, we expected to rely heavily upon project management apps and tools to organize our work and to facilitate communications. We set up Asana (a project management tool) for tasks and messaging, GitHub (a software development platform) for transferring files, and Jotform (an online form builder) for submitting collections. Over time, each of these tools became less and less useful, as task force members had to keep track of multiple web addresses for the different apps, their associated user logins, and how to use each site. Once we had developed a rhythm to our workflows, we realized that it was more effective and accessible for everyone to use email and web conferences for communications, Google Drive for document sharing, and Google Docs for tracking collections. Asana has been retained primarily for collection management and reports, and is only used by the project metadata coordinator and project lead.

Future Directions

Plans for the next two years include developing a sustainable model for service operations, growing the collection of geospatial metadata guided by the development of collection development policies and planning, leveraging expertise within the project to grow expertise in the broader GIS community through geospatial metadata outreach and education, and strategic planning to assess potential areas of strategic expansion in scope and establish our role in the larger geospatial metadata ecosystem. These goals resulted from collaborative conversations between all participating members of the project. The steering group then consulted with sponsors and stakeholders to refine the goals, as these goals will form the core of our work over the next two years.

Conclusion

Based on our experience, best practices for a successful collaboration include: strong originating documents with clearly defined roles; subgroups for specific tasks to streamline decision making; designating a project lead to keep everyone moving forward; having at least one person full-time on the project; and equally shared costs and benefits across collaborating institutions. Strong originating documents with clearly defined roles give the project lead a clear direction to follow and ensure that project participants know what is expected of them in order for the project to succeed. The creation of subgroups for specific tasks streamlines the decision making process by allowing a small, focused group to review topics and make decisions. Bringing the recommendations to the larger task force for consensus provides the opportunity for

all institutions to contribute to decisions and directions of the project. Sub-groups are opt-in, but the steering group takes special care to include members from multiple institutions in order to encourage collaboration. The idea to utilize sub-groups, and the groups themselves, evolved over time as needs were identified regarding specific topics. Task-specific groups to date include metadata, usability & interface, and collection development. While the collaboration thrives on input from all participants, designating a project lead provides an overarching vision for the project, informed and guided by a steering group and administrative project sponsors, keeps everyone in the project moving forward. Including funding for at least one full-time person, the Project Metadata Coordinator in this case, allows that person to focus on the details and guide the work of the individual participants. Working in concert with the project lead, this level of specific attention ensures that important details are not overlooked as the project develops. Lastly, equally shared costs and benefits underlies and enhances the collaborative nature of the project, as all participating institutions have an equal financial stake and an equal voice in guiding the development of the project.

Pursuing our geospatial data project as a multi-institutional collaboration allowed us to leverage our individual resources effectively for the common benefit. The BTAA Geoportal provides a discovery option to make finding geospatial data and scanned maps easier for all of our users. We also contribute best practices to the geospatial metadata ecosystem. Our flexible, iterative approach, continually revising workflows and adapting to new technologies and opportunities regarding all aspects of the project, positions us well to sustain the success of our project for the long term.

Figure 1: BTAA Institutions that are participating in the BTAA Geospatial Data Project (as of July 2017).

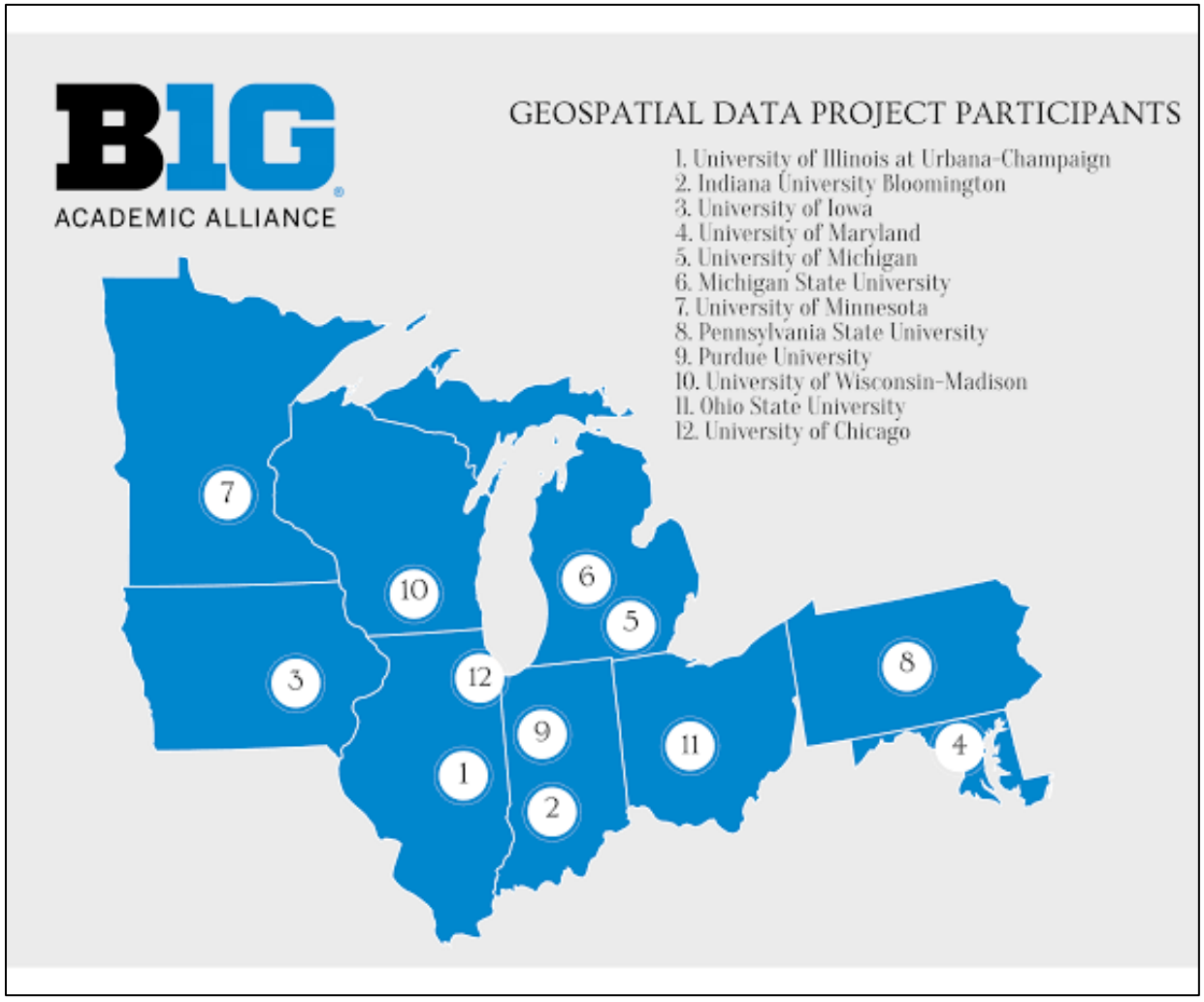


Figure 2: The Collaborative Metadata Workflow

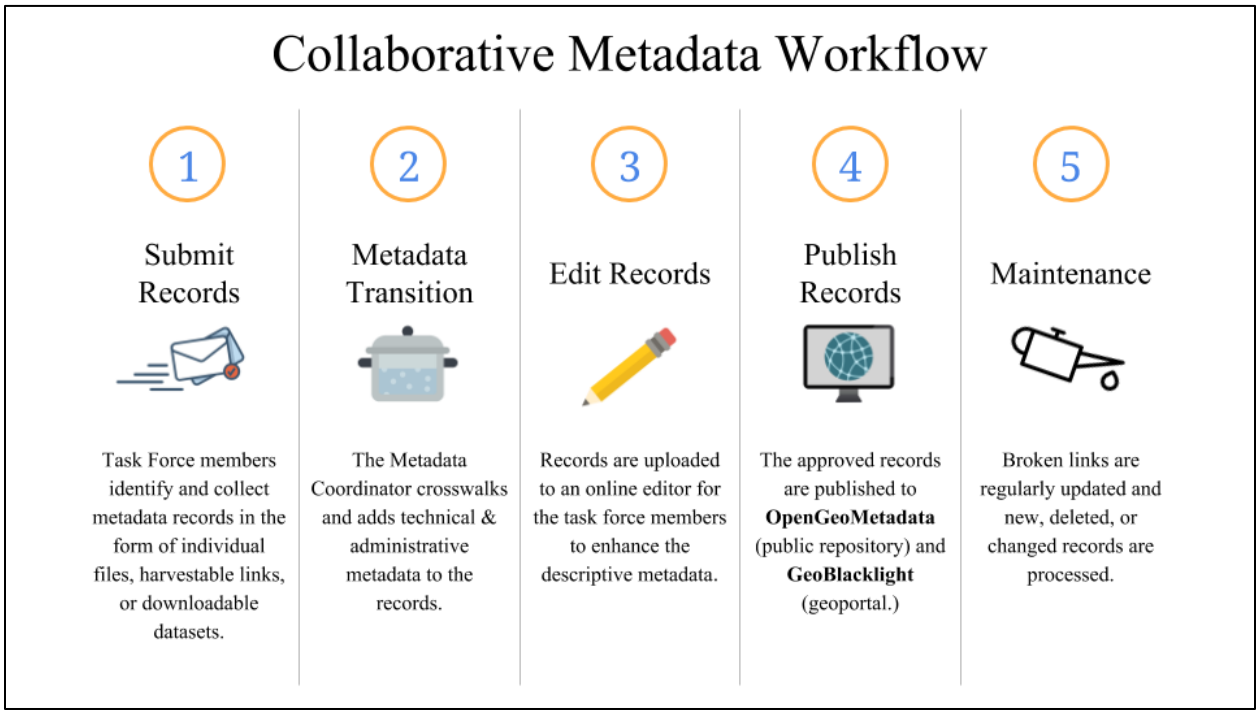


Figure 3: Original search facets, showing administrative metadata of Institution and Collection.

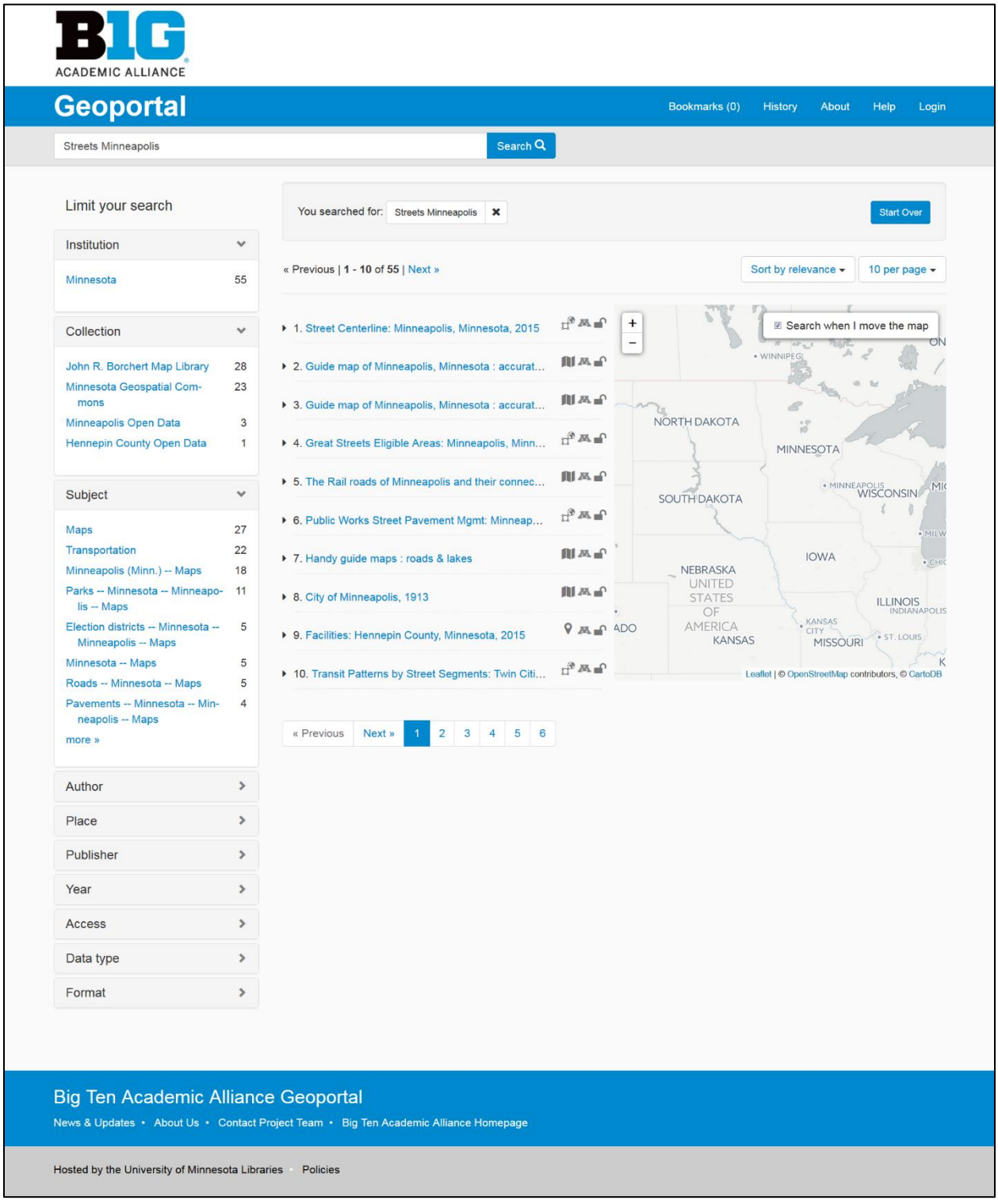
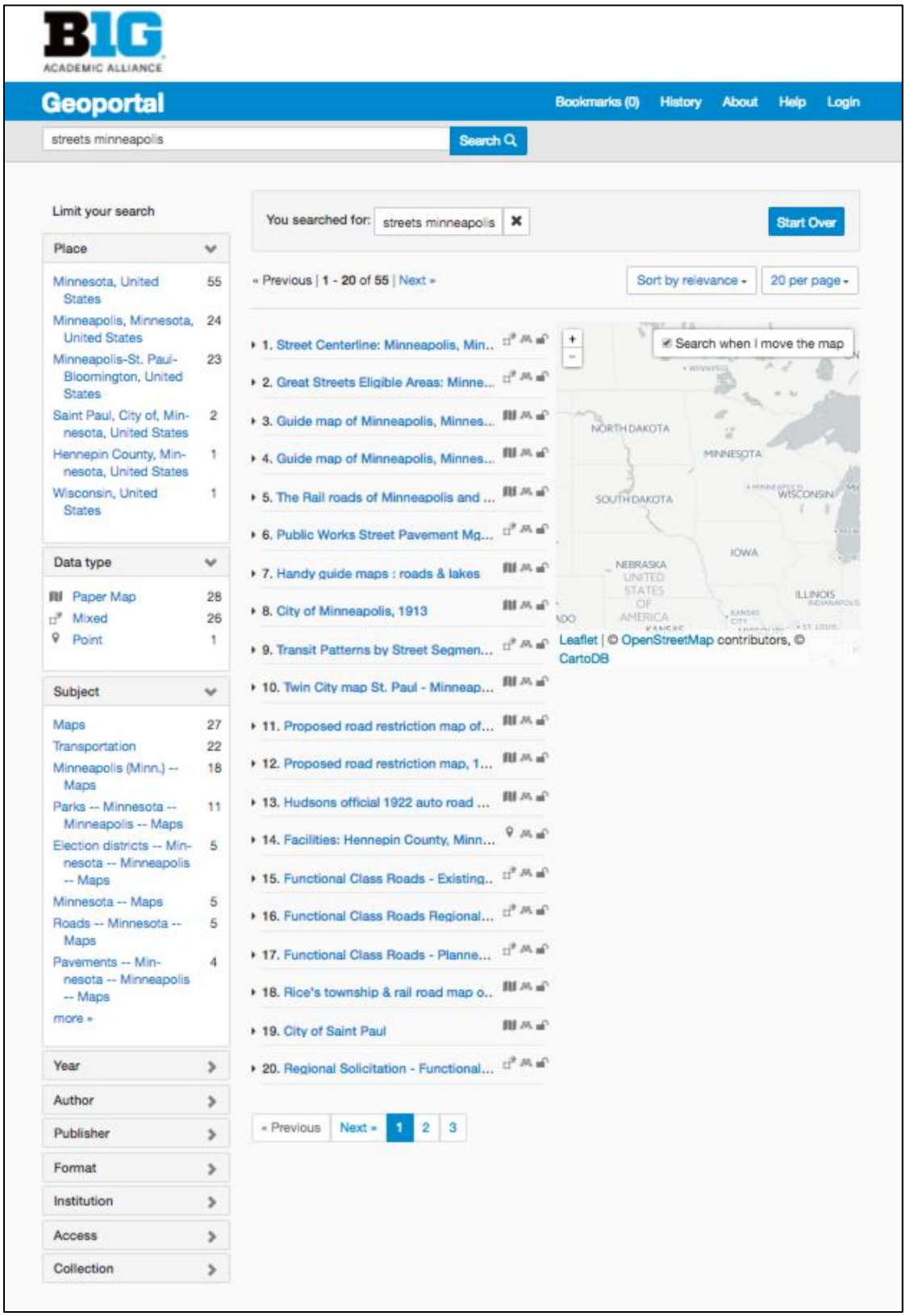


Figure 4. The user testing indicated that the top facets showing administrative metadata of Institution and Collection were not helpful for the users, so we replaced them with Place and Data Type.



[1] Big Ten Academic Alliance Geoportal, https://geo.btaa.org.

[2] Lars Bernard, Ioannis Kanellopoulos, Alessandro Annoni and Paul Smits, "The European Geoportal–– One Step Towards the Establishment of a European Spatial Data Infrastructure," *Computers, Environment and Urban Systems* 29, no. 1 (2005): 15-31. https://doi.org/10.1016/j.compenvurbsys.2004.05.009; Max Craglia and Alessandro Annoni, "INSPIRE: An Innovative Approach to the Development of Spatial Data Infrastructures in Europe." *Research and Theory inAadvancing Spatial Data Infrastructure Concepts* (2007): 93-105.

[3] Stephen Appel and Marcy Bidney. "Geodex 2.0: Saving a Legacy Map Series Cartobibliography," *e-Perimetron* 11, no. 4 (2016): 160-161, http://www.e-perimetron.org/Vol_11_4/Appel_Bidney.pdf

[4] Christine Kollen et al., "Geospatial Data Catalogs: Approaches by Academic Libraries," *Journal of Map & Geography Libraries* 9, no. 3 (2013): 276-295.

[5] Kollen et al., "Geospatial Data Catalogs: Approaches by Academic Libraries," 282-289.

[6] The OpenGeoPortal, http://opengeoportal.org/.

[7] Kollen et al., "Geospatial Data Catalogs: Approaches by Academic Libraries," 285.

[8] Florance et al., "The Open Geoportal Federation," *Journal of Map & Geography Libraries* 11, no. 3 (2015): 376-394, DOI: 10.1080/15420353.2015.1054543.

[9] Florance et al., "The Open Geoportal Federation," 379.

[10] GeoBlacklight, http://geoblacklight.org/.

[11] Earthworks, https://earthworks.stanford.edu/.

[12] GeoBlacklight, http://geoblacklight.org/.

[13] Darren Hardy and Kim Durante, "A Metadata Schema for Geospatial Resource Discovery Use Cases," *Code4lib Journal* 25 (2014): http://journal.code4lib.org/articles/9710.

[14] OpenGeoMetadata, https://github.com/OpenGeoMetadata.

[15] Lorcan Dempsey, "The Recombinant Library," *Journal of Library Administration* 39, no. 4 (2003): 116.

[16] Louise F. Spiteri, *Managing Metadata in Web-Scale Discovery Systems* (London: Facet Publishing, 2016), 18.

[17] Hussein Suleman and Edward A. Fox, "A Framework for Building Open Digital Libraries." *D-lib Magazine* 7, no. 12 (2001): 3.

[18] Carl Lagoze, Dean Krafft, Tim Cornwell, Naomi Dushay, Dean Eckstrom and John Saylor "Metadata Aggregation and "Automated Digital Libraries": A Retrospective on the NSDL Experience." *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries - JCDL 06* (2006): 233.

[19] Muriel Foulonneau and Timothy Cole, "Strategies for Reprocessing Aggregated Metadata." *Lecture Notes in Computer Science* (2005): 295-297.

[20] Digital Public Library of America, "An Introduction to the DPLA Metadata Model," (2014). https://dp.la/info/wp-content/uploads/2014/03/Intro-to-DPLA-metadata-model-2014.pdf

[21] Open Geospatial Consortium Standards and Supporting Documents, http://www.opengeospatial.org/standards/cat

[22] Comprehensive Knowledge Archive Network (CKAN), https://ckan.org/

[23] W3C Data Catalog Vocabulary (DCAT), https://www.w3.org/TR/vocab-dcat/

[24] Lagoze et al., "Metadata Aggregation and "Automated Digital Libraries": A Retrospective on the NSDL Experience." 233.

[25] M.A. Matienzo and A. Rudersdorf, "The Digital Public Library of America Ingestion Ecosystem: Lessons Learned After One Year of Large-Scale Collaborative Metadata Aggregation," *Proceedings of the International Conference on Dublin Core and Metadata Applications* (2014): 16.

[26] Matienzo and Rudersdorf, "The Digital Public Library of America Ingestion Ecosystem," 17.

[27] Bruce Godfrey and Jeremy Kenyon, "The Geospatial Metadata Manager's Toolbox: Three Techniques for Maintaining Records," *Code4lib Journal* 29 (2015). Accessed June 8, 2017. http://journal.code4lib.org/articles/10601.

[28] Foulonneau and Cole, "Strategies for Reprocessing Aggregated Metadata," 295-296.

[29] Marcy Bidney, Ryan Mattke, and Kathleen Weessies. *A Collaborative Vision for Spatial Scholarship Across the CIC* (White paper, 2012): http://z.umn.edu/cicspatial

[30] Bidney, Mattke, and Weessies, *A Collaborative Vision*, 2.

[31] Bidney, Mattke, and Weessies, *A Collaborative Vision*, 2.

[32] Bidney, Mattke, and Weessies, *A Collaborative Vision*, 6.

[33] https://z.umn.edu/cicgeoproposal

[34] Mary L. Larsgaard, *Map Librarianship: An Introduction* (Littleton, CO: Libraries Unlimited, 1998), 188.

[35] Data.gov: The Home of the U.S. Government's Open Data. http://www.data.gov

[36] GIS Inventory, https://www.gisinventory.net

[37] Project Charter, CIC Geospatial Data Discovery Project. https://z.umn.edu/cicgeoprojectcharter

[38] Task Force Member Role Description, Big Ten Academic Alliance Geospatial Data Project. https://docs.google.com/document/d/e/2PACX-1vRDcigoyN0lBjwyYbr9OOY0mrN9BlD6dbQp2YiimuRpR1dcrk-Iar1uLGLoXhPOd3MRU_bnZx1e7e7F4/pub

Appendix A: List of Participating Institutions (Year Joined)

University of Chicago (2017)

University of Illinois at Urbana--Champaign (2015)

Indiana University Bloomington (2016)

University of Iowa (2015)

University of Maryland (2015)

University of Michigan (2015)

Michigan State University (2015)

University of Minnesota (2015 - host institution)

Ohio State University (2017)

Pennsylvania State University (2015)

Purdue University (2015)

University of Wisconsin--Madison (2015)

Appendix B: List of Technologies Utilized

GeoBlacklight: geoportal platform

ArcCatalog: desktop metadata translation for GIS records

MarcEdit: desktop metadata translation for map records

Oxygen: batch editing for XML documents

GeoNetwork: online metadata editing ISO 19139 standard

Omeka: online metadata editing for Dublin Core standard

GitHub: file repository for OpenGeoMetadata

Python: batch harvesting, editing, and publishing