JournalMap: a service for geolocating academic literature and generating inferred metadata for improved discovery

NLG Program Goal and Objective: 3.2

Summary. The University of Idaho (UI) Library seeks $249,700 from the IMLS NLG program to develop JournalMap over three years. JournalMap is a prototype web service and database that indexes journal literature, parses it for the location of the study area, and then records that location. Once obtained, it uses the study locations to generate further metadata based on the location using spatial data layers such as soil type, precipitation, and biome. Using this system, we can identify environmentally context-similar papers that enhance the precision and recall of searches in the areas of agriculture and environmental research. This project builds on an FY20 planning grant LG-246411-OLS, which allowed us, with partners at Kansas State University (KSU), to plan out the requirements and design of this service. Funds will be used to develop the core application and service and integrate the system with two partner institutions, KSU and the University of Arizona (UA). In each case, partners will bring different types of document collections and database systems that require integration of the service in different ways. JournalMap will create an "Unpaywall-like" database for geolocation metadata of literature, along with the associated contextual metadata. At the end of this grant, JournalMap will be scalable, rely on automated geolocation, and be open access for use of the API. Key secondary elements of the project will be designing replicable integrations with partner's systems and a sustainability plan for JournalMap after the project.

Project Justification. Researchers have documented a problem in agricultural and environmental information-seeking behavior. Resource managers, students, policy-makers, landowners, and scientists have difficulty finding information that is salient to the context of their work (McNie 2007; Wallis et al. 2011; Schmitt and Butler 2012; Karl et al. 2013). While access to scholarly literature has become dramatically easier in recent years, it can be difficult to sift relevant documents out of search results. Relevance is determined, in part, by the context of the work, and for many topics, context can be determined from location (Karl et al. 2013). In much of science, the results (or even the questions asked) are influenced by the place and time in which it was conducted (Livingstone 2003). However, existing bibliographic search tools (e.g., Google Scholar, Web of Science, library catalogs) still focus primarily on the *what* of research while largely ignoring the *where.* This prevents efficient searching based on research location or on location-related attributes including environmental, social, and economic features (Karl et al. 2013). Thus, the lack of usable location information and the corresponding lack of location-based search tools limits knowledge discovery.

The value of georeferenced literature databases has been established in many fields including ecology and conservation (Page et al. 2011, Martin et al. 2012), land management (Wallis et al. 2011), environmental science (Schmill et al. 2014), and infectious disease (Hendrickx et al. 2010). However, in most cases, assembly of these databases is a laborious process of manually geotagging articles and are not typically sustainable or efficient. Additionally, until a very large number of articles are georeferenced, the value of geographic-based literature searching is very limited. Automation is a logical step towards scalability, but, has seen limited use due to the complexity of toponyms and their identification (Gritta et al, 2017; Gritta et al, 2019). Named-entity recognition software, like Spacy, combined with geocoding resources, like Geonames.org, have introduced the capacity to do automated geographic indexing more effectively.

The NLG program has focused objective 3.2 on innovative approaches to collection management and on improving cataloging and inventory processes. JournalMap offers an automated approach to classifying articles by geographical location (Karl 2018, Kenyon et al. 2021), an important step toward scaling up and making feasible geographic-based searching. JournalMap's approach allows for a single, open service in which to collectively compile georeferenced location metadata along with their inferred location-related attributes.

Project Work Plan. In the first year, we will develop our redesigned JournalMap service into a production-level service and application (note: the current website for JournalMap does not feature this redesign). During our FY2021 planning grant, we re-engineered the original JournalMap from Ruby to Python, and designed the database to include any document type (originally restricted to articles) using the Citation Style Language (CSL) schema and to accommodate any type of spatial object (originally restricted to coordinate pairs, or points) using GeoJSON format. Our goal in year one will be to ensure that the service is ready to receive new collections through its provider API and return geolocated and enhanced metadata about indexed resources. From our FY20 planning grant, the base functionality and database design is already in place, includes a interface allowing for display of a range of spatial

object types and a provider dashboard for editing and modifying records. Implementation will require developer effort to bring these elements up to a production-level standard, test it rigorously, and make it ready for public use.

Also in year one, we will continue refining our approach to automated geoparsing. The tools to do this perfectly do not yet exist. However, we have implemented the Mordecai application, which uses the aforementioned Spacy and Geonames tools, to parse and geolocate items (Halterman 2017). We have ongoing work in 2021-22 that will advance our use of Mordecai to select for more precise information and assign a confidence interval based on the accuracy of the document. If awarded the implementation grant, we will be able to implement this system in year one.

As a strategy for building out our content and expanding the tool's features with other systems, UI will engage with several partner institutions. KSU will complete integration of JournalMap and their Croplands Research Database as well as their natural history database, Biodis (currently offline). UA will focus on integrating the journal archives for *Rangeland Ecology and Management* and their discovery tool, the *Rangelands Gateway*. Our partners will begin in year one to develop pipelines for moving content into and out of JournalMap. KSU has already developed an approach that uses Zotero for moving document metadata, but UA will develop a pipeline for Dspace repository metadata, and others. Third, we will convene an advisory group designed to evaluate our progress, provide critiques, and assist with developing contacts and devising a sustainability plan. Advisors that we expect to serve include researchers with the USDA Agricultural Research Service, technical services librarians, and experts in natural language processing.

The second year will complete any remaining core development for the service. We will move towards a strategy of developing feature requests from our partners, as well as tracking and eliminating bugs. Our goal in year two is to do at least one round of updates and enhancements. As our partnerships continue, we will add new spatial data layers and generate filters relevant to their content areas. For example, the Rangelands Gateway contains literature on social and economic dimensions of agriculture, rather than just biophysical dimensions. Adding relevant data layers for these subjects will improve not only the relevance of the tool for those partners, but for other users in other domains as well. We will also continue to ingest articles, focusing first on open-access literature, for the sake of expediency. Finally, a year two goal will be to make progress in developing a sustainable system for maintaining the service. Our advisory group will assist in providing feedback and providing connections to others that have experience with these issues. JournalMap might work on a freemium plan, based on potential user features, but that and other options will be explored. Our goal, however, is that the automated geolocation of literature, the assignment of metadata, and the capacity to use the API to add and return data will always be available to non-commercial entities.

In year three, we plan to continue to the process of feature enhancement and bug fixing as necessary. However, at this point, we plan to be present at numerous library and scientific society conferences presenting JournalMap and showcasing how it can be used. Secondly, we aim in year three for an evaluation element through our advisory group. Based on our goals and outcomes, we will evaluate (1) the quality of the application/service, including the accuracy of the automation in selecting the right location, (2) the range and diversity of the collections, and (3) the precision of attribute data derived from the locations.

Diversity Plan. JournalMap is an open-source web service that offers its metadata for re-use. In essence, the system will not restrict access to anyone, enabling access at all educational levels. Further, we plan to explicitly build out representation of content from non-Western journals and repositories such as African Journals Online (AJOL) journals and our partners, KSU and UA, have content collections heavily representing studies done in the rural parts of the United States, including Western rangelands and in the landscapes of indigenous communities. In its user interface, JournalMap equalizes representation globally by focusing on the contextual similarity of study sites in its results.

Project Results. The project will have several results. First, JournalMap will be available as a production-level service, with an API for adding material and retrieving it. Second, JournalMap will be available as a customizable web interface, the primary feature is the limitation to a provider's collection. We anticipate sharing the system at national meetings to encourage use and this will produce presentations and papers detailing the development and aims of the service. Lastly, each of our partners will end up with JournalMap integrations into their systems.

Budget Summary. The University of Idaho requests $249,700 for the project. The core expense is development time by our personnel, which is estimated at $65,000 for the three years. UI project team, including principal investigators, will require $33,000 over the time period. We have also included approximately $10,000 for project travel over the three years, primarily for conferences. UI indirect costs for this project will be 38%, which will come to $41,040 of the above estimates. $40,500 is requested for KSU and $41,160 for UA to participate, which includes their development personnel, project investigators, and local indirect rates, and creates subaward processing costs for UI of $19,000.