Planning the Creation of a Multi-Purpose, Public Facing Academic Research Database Platform for Agricultural Data with Geolocation Data Correlation

Introduction

We are proposing a web application and service to improve search and discovery of agricultural and environmental literature through geographic searching. This service will be built as an evolution of two already successful applications: Global Research Alliance (GRA) on Agricultural Greenhouse Gases' Croplands Research Database (CRDB, see URL list in supplemental material) which is maintained by Kansas State University Libraries and JournalMap. The CRDB provides researchers with a curated database of literature about climate change effects on crop agriculture. JournalMap provides a model for geographic literature searching (Karl et al. 2013) and a powerful mechanism for extracting geolocation data from academic papers and applying it to articles as a metadata field enabling location-based search (Karl 2018). Our proposed system is designed as a modular, flexible, multi-modal, open source web service for integrating with databases, catalogs, or discovery systems used in library and publishing environments. In this planning grant, we will focus on agricultural and environmental scientific literature, but the proposed concept is extensible to any knowledge domain and resource type for which location is a relevant search parameter.

IMLS has a stated goal of increasing public access to information through investment in tools and technologies that enable people of all backgrounds to discover library collections and resources. Our project fits this objective in part because the lead institutions are land-grant public universities. In both Idaho and Kansas, our libraries are dedicated to supporting the educational and economic aims of our states, and our institutions, and see an opportunity to make agricultural and environmental information more accessible by enhancing existing methods of search and discovery. We find that by doing so, we support not only members of our universities, but through leveraging open, public deliverables, we support government officials, citizens, farmers, ranchers, and the public, regardless of background or identity. For this planning grant, our primary audience will include researchers who need effective tools for literature searching, and libraries and data service providers who facilitate searching and discovery.

Statement of National Need

While access to scholarly literature has become dramatically easier in recent years, 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, climatic, social, and economic features (Karl et al. 2013). Yet much of the scholarly literature is either location-based, its applicability dependent on spatial context, or the results (or even the questions asked) influenced by the place and time in which it was conducted (Livingstone 2003). Thus, the lack of usable location information from literature and the corresponding lack of location-based literature search tools limits knowledge discovery.

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, 2011; Schmitt and Butler 2012; Karl et al. 2013). Salience, in this case, means information that is not only topically relevant but possesses spatial or temporal attributes aligned with the user's information need (Figure 1, Karl et al. 2013). For example, if a researcher or land manager needs information on practices to control soil erosion in Namibia, a topical literature search will include many sources that are not appropriate to the region or its soil types. Locally-generated research would be difficult to find without well-developed social networks or *a priori* knowledge (see Zimmerman 2007), and discovery of relevant information from other regions with similar soils and climate (e.g., southern New Mexico) would be near impossible without incredibly broad subject knowledge.

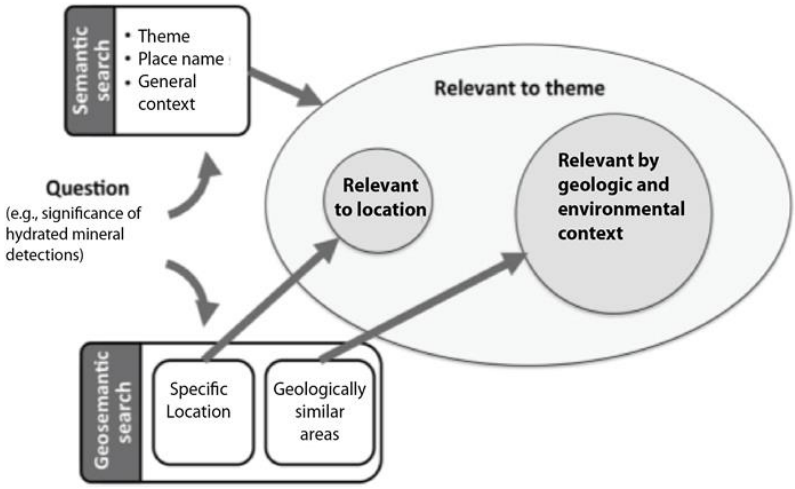

Figure 1. The ability to search by theme and location (i.e., geosemantic search) can improve the relevance of search results. Figure from Karl et al. (2013).

Most literature database tools have been constructed without an appreciation of the complexity of spatial searching, relying largely on geographic place names to describe locations. This approach, however, is flawed because there are no publication standards for reporting place names. For example, Karl (2018) described a Web of Science search of ecological literature for studies with "Chihuahuan Desert" in the abstract and associated indexed information. Of the more than 800 articles returned by this search, only one third of them actually occurred within the Chihuahuan Desert (Figure 2a), due largely to the presence of irrelevant place names in articles (see Karl 2018). Additionally, many more studies occur within this region but used different names to describe their study areas (Figure 2b).

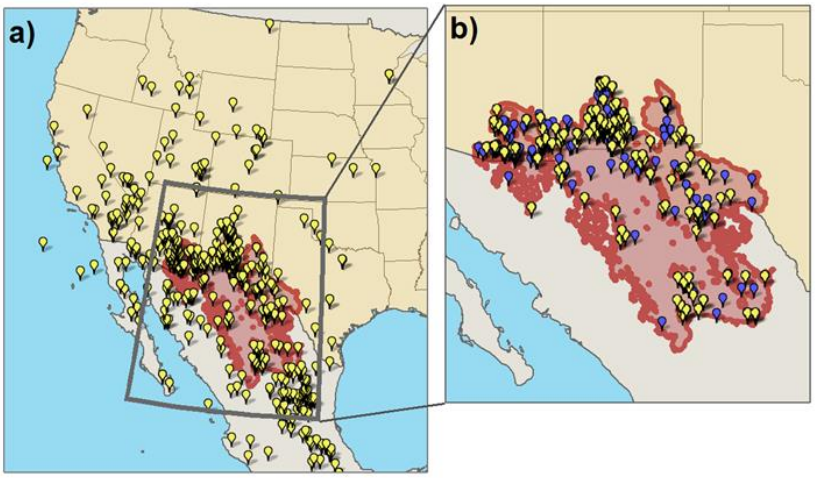

Figure 2. Example of the challenge of searching for scientific literature based on location using existing search tools. Only 33% of over 800 articles returned from a Web of Science search for the location name "Chihuahuan Desert" were within that area (red shaded region, Map a). A search of JournalMap.org returned many additional articles in the area that did not use the term "Chihuahuan Desert" (Map b). From Karl (2018).

Several search tools have been developed that have begun to change the thematic-only literature search paradigm through map-based searching. For example, JournalMap (Karl et al., 2013) provides a map-based search interface for georeferenced journal articles and an API for embedding search results and article maps.

The USGS Science Base website permits basic map browsing of USGS reports and articles by agency scientists from geographic coordinates or bounding boxes assigned to each source. The data repositories such as Pangaea, Earthworks, and DataOne map locations of datasets, many of which can be tied back to published articles. Article abstracting services like GeoRef and CAB Abstracts assign general geographic regions (i.e., place names) to articles. BioStor extracts geographic locations from historic articles in the Biodiversity Heritage Library and provides article-level maps of specimen locations (Page, 2011). There are also numerous examples of "georeferenced bibliographies" that offer maps of article locations related to specific themes (e.g., van Vliet et al. 2012, Pert et al. 2015, Howell et al. 2019). Of the existing examples, most assign locations to published articles either via their source data (e.g. Pangaea), self-reporting by the authors or manually (e.g. ScienceBase, CAB Abstracts). Currently, only JournalMap and BioStor employ automated geolocating algorithms to mine location information (from printed geographic coordinates) from article text. However, these approaches will continue to be of limited utility until the total amount of georeferenced literature is greatly increased, and this will hinge on developing new techniques for rapidly and accurately georeferencing existing literature.

Offering searchable maps of literature based on location or geographic names is a step in the right direction, but most of the available services do not provide for identifying literature from contextually-similar but geographically-separated regions (see Namibia erosion example above). Of the sites offering geographic searching for literature, only JournalMap (but with limited spatial attributes; Karl et al. 2013) and the GLOBE project (focused solely on land-use change studies; Schmill et al. 2014) provide functionality for searching based on location similarity. Enabling similarity-based searching is possible through inclusion of additional and readily available environmental (e.g., elevation, biome), social/political (population, political regime), or economic (e.g., GDP) spatial layers.

The value of georeferenced literature databases has been established in many fields including ecology and conservation (Page 2011, Martin et al. 2012), land management (Wallis et al. 2011, Karl et al. 2013), environmental science (Schmill et al. 2014), library sciences (Johnson et al. 2009, McKee 2019), and infectious disease (Hendrickx et al. 2010). However, in most cases, assembly of these databases is a laborious process of manually geotagging articles and ultimately these efforts are typically not sustainable beyond their original research objectives (see supporting document "Selected efforts similar to JournalMap"). Additionally, until a very large number of articles are georeferenced, the potential for geographic-based literature searching is very limited. JournalMap's automated approach to identifying and extracting geographic coordinates from articles (Karl 2018) is an important step toward scaling up geographic-based searching, but articles with coordinates account for only about half of papers published in ecology journals (and less for other knowledge domains; Karl et al. 2018). Thus, in order to sustainably scale-up the concept of geographic literature searching, more robust, automated georeferencing approaches need to be developed and implemented.

Integration with other tools/efforts

JournalMap demonstrates the capacity to search for a given topic and use its location to find other sources from areas with similar attributes. But the greater value of JournalMap would be integrating it with other search tools and databases to improve knowledge discovery and infer relationships with other resources (e.g., literature, images, datasets), especially those based in libraries. JournalMap was originally developed as a web service, but not in coordination with other efforts. Thus, integration is challenging. Once an article is georeferenced, its accompanying metadata becomes richer over time with no additional labor cost as that location is overlain with new data layers. In fact, adding more data layers and new attributes to an indexed article is a trivial additional cost. A revised JournalMap designed to provide geographic search capabilities and locational metadata to existing literature databases or data repositories, in a constellation of different subjects, would allow institutions to stand up a powerful, customizable database solution to provide maximum benefit to their users.

The CRDB as a Test Case

The CRDB was developed to support the GRA's mission of reducing greenhouse gas intensity and improving production efficiency of cropland systems. The CRDB was originally created nearly 10 years ago at the request of the GRA's croplands research group. The current CRDB, developed and maintained by Kansas State University Libraries, provides enhanced metadata for over 7,500 scientific articles related to crop agriculture in different climate systems around the world. The GRA sought a controlled vocabulary applied to citations that would assign a crop type (e.g., corn) or cropping system (e.g., irrigated crops), climate regime, and country to each article. As a location attribute, though, country is often unhelpful as many countries span myriad environmental, climate, and sociodemographic zones. Access to the knowledge in the CRDB by the 60+ countries in the GRA would improve dramatically if more precise location information were associated with the articles and if the articles were discoverable via robust map-based searches. A truly functional geographic literature search engine, populated with a robust corpus of literature, would also facilitate identification of new articles for the CRDB.

Functionally, the CRDB also exhibits basic design features common to many agricultural and environmental databases. It is primarily an enhanced bibliography, represented as a web application with a keyword-driven search and a series of facets to filter the results. These databases are common. For example, the University of Minnesota maintains AgEcon Search, which provides a similar enhanced bibliographic search to agricultural economics papers. In both cases, the functionality could be improved further by integrating JournalMap's automated indexing and spatial metadata. The CRDB thus provides a test case for designing an integration for one straightforward bibliographic resource, and an example of how others might follow suit.

Project Design

Our intention with this IMLS planning grant is to investigate, design, and prototype improvements to and integrations between JournalMap and the CRDB. Longer term, we will use the redesigned system as a model to begin scaling the project with other partners. Our project has three goals for the one-year period of performance. First, we will redesign the underlying (i.e., back-end) architecture of JournalMap to more easily support the automation of article capture, classification, georeferencing and integration of CRDB content. We will import the CRDB current bibliography into JournalMap as a "collection" - one of the ways of organizing literature in JournalMap – and implement enhancements to JournalMap's geotagging and classification algorithms to better serve the needs of the CRDB. We will then serve CRDB content through a prototype JournalMap interface and via web services to illustrate the ability to integrate content in JournalMap from other sources as well as embed JournalMap's geographic searching into third-party bibliographic applications like the CRDB.

Our second goal is to use the new JournalMap system architecture and prototype interface to begin discussing functional requirements for JournalMap/literature-database integration with partners. This will include discussions with colleagues at the Global Biodiversity Information Facility (GBIF), the National Agriculture Library, USDA Agricultural Research Service, North Carolina State University, the University of Arizona, the University of Minnesota, and possibly others. During this period, we will present our work at library conferences, such as the Digital Library Federation, to demonstrate the developments of the project and solicit participation from interested colleagues. This will prepare us for our third goal, developing an IMLS National Leadership Grant project grant proposal to be submitted at the conclusion of the year to begin scaling the system designed in this planning grant phase.

Goal 1. Create an application architecture and prototype for JournalMap and the CRDB.

Objective 1: Redesign existing architectures for flexibility and integration. JournalMap is built on technology that is nearly a decade old and does not permit the flexibility to easily integrate with other bibliographic

projects. While functional, JournalMap's underlying database code base needs to rebuilt to be more extensible and scalable. For example, in its current iteration, JournalMap is written in Ruby on Rails, a web framework that has seen declining use and is not optimal for advanced applications like machine learning (Gonzales 2019; Mista 2019). Second, JournalMap uses ElasticSearch on top of a custom relational database to provide efficiency, but this constrains JournalMap to using only points to represent article locations (as opposed to polygons), precludes the capacity to conduct spatial queries, and limits easy integration with other literature databases or applications (e.g., citation managers). In both cases, other web frameworks and database designs are more suited to sustainable development, and advancements in geospatial infrastructure could enable an equally efficient search, while enabling new capabilities.

The CRDB is not configured for automated ingestion or use outside of its existing search interface. Thus, it cannot be integrated easily with JournalMap in its current form. Improvements to the CRDB including automating article import, assignment of a controlled vocabulary for cropping system (using crop system spatial data layers, like those found in USDA's CropScape), and geotagging articles will enhance the value of the CRDB. Redeploying the CRDB via JournalMap will improve its impact and offer a good example of the value of integrating JournalMap with existing bibliographic databases. Integrating the CRDB into JournalMap will eliminate the need for assigning country and climate classification manually and enable the inclusion of many other environmental, climate, social, and economic search attributes important to crop agriculture.

Our plan for Objective 1 is to work with computer science experts at the University of Idaho's Northwest Knowledge Network to design a new architecture for JournalMap and identify the tools and technologies necessary to fully implement that design under a future project grant. The new architecture for JournalMap will prioritize use of existing bibliographic database schema (e.g., Dublin Core Metadata Initiative), best practices for storage and processing of spatial data. The project team will meet regularly during the fall of 2021 to outline the current state of JournalMap and the CRDB and begin discussing optimal approaches to redesign and integration. We will invite feedback quarterly from our stakeholder's advisory group on feasibility, extensibility, scalability of new design options beyond our two institutions, as well as other comments regarding the system requirements. Our primary output through Objective 1 is a series of design documents, system architecture diagrams, database schemas and relationships, and wireframes of a prototype interface.

This objective bears few risks as our primary goal is design and prototyping, rather than production software development. That said, one potential risk we recognize is the potential for the architecture to be too specific to CRDB. To mitigate this, we will continue to revise as we proceed through the second half of the grant period and make modifications in response to our advisors' interests, criticisms, and specifications. For this objective, all key project personnel will participate in the design process.

Objective 2. Test new approaches to geotagging and spatial infrastructure. A number of tools have been developed since 2010 which can improve and make feasible automated location extraction. Currently, JournalMap uses a straightforward, but fairly restrictive, parsing of geographic coordinate variants to determine article locations automatically, as other natural language processing approaches had not been integrated into JournalMap (see Karl, 2018). Articles not automatically georeferenced in this manner (i.e., do not have geographic coordinates) go into a queue to have locations assigned manually. This creates a significant bottleneck to adding content to JournalMap. Latitude and longitude coordinates found in articles are intersected with a series of spatial data layers to determine the environmental attributes of the location, all of which are then stored in a relational database for efficient search and retrieval.

We will explore natural language processing techniques beyond coordinate parsing to improve the scalability of JournalMap's location extraction algorithm. Specifically, we will test: 1) the Stanford CoreNLP (Natural

Language Processing) resource distributed through Stanza, 2) spaCy, 3) and NLTK. Each of these software libraries performs well in named-entity recognition (NER) by using statistical models of various languages, including English, to determine the likelihood that a word is a noun and/or the name of entity, such as a place, organization, or person (Harrington 2019, Schmitt, et al 2019, Stahlman et al. 2019). We will also test tools available to us through the University of Idaho's ESRI site license, such as LocateXT. LocateXT is an ESRI extension that performs NER on unstructured text such as journal articles or abstracts and geocodes those items based on a source document of names and associated places (ESRI, 2020). We will use existing place-name resources, such as Geonames.org or the Geographic Names Information System (GNIS) to provide the name-coordinate pairing. To reduce the effect of extraneous place names on geotagging accuracy, we will develop rulesets to rate likelihood that detected place names describe the study location using attributes including location of the place name in the article, semantic context, and filtering of common species names (e.g., "Canada Goose").

During the fall of 2021, our plan for testing will start with manually geotagging a representational set of articles to create a benchmark dataset. We currently have access to full-text versions of all literature in the database. We will then test each of the four approaches: LocateXT, spaCy, NLTK, and Stanza, to determine which leads to the best NER for geographic names in our articles. Then, we will develop conditions for reducing the retrieval of irrelevant place names. Conditions might include weighting location terms based on the section of the article in which they appear, or checking for clusters of closely situated different place names that suggest a flurry of mentions, not a representative study area. Results of these algorithms will be summarized and visualized for comparison before deciding on a final outcome. Natural language processing has been explored by both Dr. Jason Karl and Jeremy Kenyon in other projects, and they will lead the exploration and testing of including these tools. Bruce Godfrey, GIS Librarian, at the University of Idaho Library will provide the expertise and experience integrating the ESRI tool and geographic names services into our design. Additionally, we will seek feedback from project partners on acceptable location error rates and how to design a system that is robust to location errors and provides for their identification and correction. Advisors will be introduced to our progress and given the opportunity to provide feedback or contribute as they are interested in doing so.

We perceive two potential risks for this Objective. First, the use of proprietary tools (i.e., ESRI's LocateXT) could limit reuse of geolocating algorithms developed for JournalMap. The project team has previously produced numerous resources for public use and consumption using ESRI technology to no ill effect or restriction. However, we plan to ensure that any software we design is not entirely contingent on ESRI licenses. For example, indexing using ESRI technology will not affect any other database using the JournalMap bibliographic index. Most of the data processing can be done using the free API and using Open Geospatial Consortium's (OGC) set of standards and services.

A second notable risk to the long-term success of this project is that automated location-parsing algorithms will return erroneous or spurious results, thereby compromising the basic premise of geographic-based literature searching. This is driven by the fact that articles are often filled with place name information not related to the study area in question. Thus, developing an approach that is able to reduce the error rate of irrelevant named entities is a key priority. This is the general classification problem that poses a significant challenge but also creates an opportunity for fundamental innovation, and as noted earlier, allows us to scale a heretofore little-used approach to context-similar searching and indexing.

Objective 3. Developing a Prototype for JournalMap & CRDB. We will develop a prototype of JournalMap using the new architecture and technology (e.g., web framework) designed in Objective 1 that will include an application search interface (website) and web services (API). The new prototype will focus on streamlining article ingestion, implementing improved geotagging algorithims from Objective 2, adding autoclassification from a controlled vocabulary, and improved geographic searching (e.g., adding more layers for search filtering). With the improved backend and API of JournalMap, we will also develop a separate demonstration interface for the CRDB that pulls data directly from JournalMap via the API to showcase the potential for JournalMap integration with other bibliography efforts. Livia Olsen and IT staff from Kansas State University Libraries will contribute to this objective, along with Jennifer Hinds, a web application designer from the University of Idaho with expertise in developing spatially-enabled web applications. Keeping in mind inclusive user experiences, we will commit to using Universal Design standards as well as using the Web Content Accessibility Guidelines (WCAG) 2.x to ensure maximal web accessibility for this prototype. We will seek feedback from our advisors during the quarterly meetings as this process progresses, incorporating their feedback into the design specifications. Risks for Objective 3 include specifying too large an initial feature set which could result in failure to complete the prototypes. We will mitigate this risk through use of Agile-style development and prioritization of features essential for meeting the objectives of our Planning Grant.

Goal 2. Enlist Stakeholders for a Project Grant

Objective 4. *Engage with stakeholders during design and prototyping.* Each quarter, we will convene with our stakeholders' advisory group to update them and to receive recommendations and feedback on further developments. During the first several months, we anticipate producing initial design diagrams and performing early tests to improve the indexing process, while beginning to work on the user interface for CRDB. Our expectation is that the stakeholders' group will be able to provide feedback that furthers the development while not requiring their frequent involvement in the process. One risk to this activity is the potential that our advisory group is not representative of everyone who might find the tool useful. Regardless, we feel that an informed, interested group should be able to provide substantial feedback to get us closer to a widely useful design. Engagement activities will include contacting stakeholders to determine their interest making recommendations about new developments in JournalMap and the CRDB; to find out if they have ideas about other potential stakeholders; and conducting an informal survey of their expectations and requirements for these two applications.

The primary stakeholder for the CRDB is the GRA Croplands Research Group. Dr. Chuck Rice (K-State agronomy professor) and Dr. Mark Liebig (USDA ARS soil scientist) are both involved in the Croplands Research Group and will serve as advisors, giving feedback and encouraging others in the GRA to engage with this process. They are willing to present about this project at the Croplands Research Group annual meeting in fall 2020 and promote the survey (objective 5) about the CRDB and JournalMap. They will help us work with the GRA headquarters in New Zealand to promote this work and a survey through email blasts and promotion on the GRA website. This research group is only one of four research groups in the GRA so there are engagement opportunities for other disciplines within agriculture through working with the GRA.

Other advisors include Julie Kelly of the University of Minnesota Libraries, who operates the AgEcon Search Database and is interested in looking at ways JournalMap could integrate with their system along the same lines as the CRDB. Robert Olendorf of the Natural Resources Library at North Carolina State University has also agreed to serve as an advisor. He will provide feedback on our progress, having served on user experience groups with the DataOne project, among others. Jeanne Pfander, of the University of Arizona Libraries, will

also serve as an advisor. Her experience working with agricultural researchers, students, and conducting outreach to academics and management professionals and others in the U.S. Southwest, will bring expertise regarding search interfaces and the user experience. Dr. Jeff Herrick (USDA ARS soil scientist), lead of the LandPKS project, will provide exerience in developing crowd-sourced knowledge systems. Dr. Jeffrey Cambell of the USDA National Agriculture Library's Knowledge Services Division will also serve as a project advisor.

Objective 5. Conduct needs assessment and requirements for stakeholders' integrations. Beginning in Feb 2021, we will launch a survey of potential interested users, including librarians at our respective and nearby institutions, members of the GRA members of the U.S. Agricultural Information Network, and potentially others, to identify appropriate features for our interface and the potential for integration with some major existing tools. We will use the University of Idaho's Qualtrics software to issue the survey and do so under the auspices of the University of Idaho's Institutional Review Board. We will not ask any sensitive questions or gather personally identifiable information. This survey will complement our advisory group's feedback gained through meeting during the fall. We estimate that a range of integration options will need to be developed in order to make JournalMap effective to the widest possible range of partners. Many libraries do not have the resources to easily incorporate JournalMap into their systems

Identifying these key tools will give us a sense of the scale possible within the library community. For example, a potential integration with the Primo discovery layer used by approximately 30% of the academic libraries in the U.S. - including the University of Idaho and Kansas State University – would enable a wide range of libraries to use JournalMap's location-based search features within their catalog (Breeding, 2020). Examples of integration might be an embedded, searchable map feature within a record that has been indexed by JournalMap, or a recommendation tool that provides a list of articles that represent an area with a similar environmental context as the selected record. To integrate, compatible widgets built as Angular Javascript functions would need to be made so that libraries could connect JournalMap within their institutional Primo instances, and vice versa. However, this will not work for efforts like GBIF, the Rangelands Gateway at the University of Arizona, or USDA Agricola; each operates a unique and custom system. Thus, a survey will help us begin to see the possible systems where integrating JournalMap yields the greatest benefit.

Goal 3. National Leadership Grant (NLG) - Project Grant Proposal

Objective 6. Submit an NLG project grant proposal. We have discussed the potential of a future project with our stakeholders, but we cannot guarantee all will be available for a future project. Therefore, working with available and interested project partners, we will develop an NLG Project grant proposal using our new design for the JournalMap system, and for the CRDB application and set of integrations for project partners' systems to be produced between June and August 2021.

Evaluation and Measures of Success

Our Goal One efforts will be successful if we can produce an acceptable, basic architecture that serves the interests of multiple partners. A second measure of success will be that at least two partners seek to replicate the CRDB approach and implement a similar database from our design on an existing simple bibliographic resource. A third measure will be the error rate for automated geocoding. While we aim to achieve a benchmark of 95% accuracy, we will assume a 90% or better accuracy rate to be acceptable. Further, we will test our prototype user interface assessment with partners through qualitative testing and informal surveys. Goal Two efforts will be deemed a success by identifying and including these partners in the future project proposal. Goal Three efforts will be successful as we produce and deliver a proposal by September 2021.

Diversity Plan

The proposed tools for geographic literature searching contribute to a more equitable system of knowledge discovery worldwide. For example, a North-South digital divide in scholarly publishing is well described wherein the vast majority of scientific literature is generated from developed countries in the Northern Hemisphere and emphasis is placed on facilitating flow of information from North to South as a means of development aid (Britz and Lor 2007, Chan et al. 2011). Significant challenges for researchers from developing countries are getting their work published in recognized international journals (Chan et al. 2011) or having works published in local or regional journals be discoverable (Czerniewicz and Wiens 2013). This perpetuates the view that little research originates in developing countries and widens this scientific digital divide. Map-based and geographic-similarity-based literature searching could help narrow the gap for developing countries (i.e. improve South-North information flow) by increasing discoverability of research. From the example above, researchers in New Mexico could find value in Namibian soil erosion studies, but with current topical search technologies would be unlikely to find them (and for the same reasons Namibian researchers wouldn't find New Mexico research). Additionally, map-based searching could also help bring together scientific and mapped traditional knowledge (e.g., Australian Indigenous Biocultural Knowledge database; Pert et al. 2015) to increase awareness and understanding of and opportunities for marginalized cultures.

Our planning project indirectly supports the efforts of the GRA in sharing knowledge of agricultural practices to reduce greenhouse gas emissions across its 61 member countries which are from six continents and multiple Pacific islands. There are four research groups within the GRA, croplands, livestock, paddy rice, and integrative. While the CRDB is the focus of this proposal, because of the diverse research interests within the GRA, there are opportunities to reach out and discover the needs of other disciplines which could be integrated into JournalMap. Additionally, we will seek members of our project stakeholders' group who represent underserved communities or developing countries. It is our goal to create a system that contributes to the Global Knowledge Commons (Chen et al. 2011) through improved discoverability of relevant information to solve global sustainability challenges.

National Impact

This planning project is calibrated to produce three deliverables (application architecture, the prototyped interface, and the project proposal) which will be used to request funds for the next phase of this project. This project is primarily operating in the exploratory phase of maturity (Matthews, 2018). The CRDB and JournalMap have both have both proven their core concepts. Interested parties have used or continue to use the resources as they are. However, we are mindful that there are numerous improvements in the underlying technology that require us to revisit them. Thus, to quote IMLS Director Matthews (2018), we are trying to "adapt [our] recipe to a new flavor combination." In other words, we know the current approaches work. But we do not know how well they can scale; in fact, we are fairly certain that they will not scale efficiently in their current forms – in both cases, too much labor is required. Our goal in this process is to explore improvements and design a pilot. The national impact will rest first on our success at creating a robust modernized system and second, at enlisting our stakeholders to integrate the resources with their own favored tools. Based on our experiences and discussions with colleagues – both seem promising.

Disseminating context similar-searching can potentially have a profound impact far beyond the domains of agriculture and ecology (Figure 3). Currently, information retrieval in libraries tends to rely heavily on library- and publisher-generated metadata, most of which is bibliographic in nature, or utilizes broad, directly referential classifications, like a subject heading. Further, most of that assignment is done by people, or at best, might potentially be done through automated methods like topic modeling. No one seems to have added the logical step of geographic inference, perhaps because that expertise tends not to be present in libraries' metadata and bibliographic control departments. The impact of this tool can be significant, if only because if realized, it offers

an efficient, inexpensive method for providing context to search, without requiring the labor force, time, or resources to do in other ways. The cost-benefit of JournalMap is that one only indexes an article once. Once that is done, anyone can re-use the information by accessing it over an open, free API.

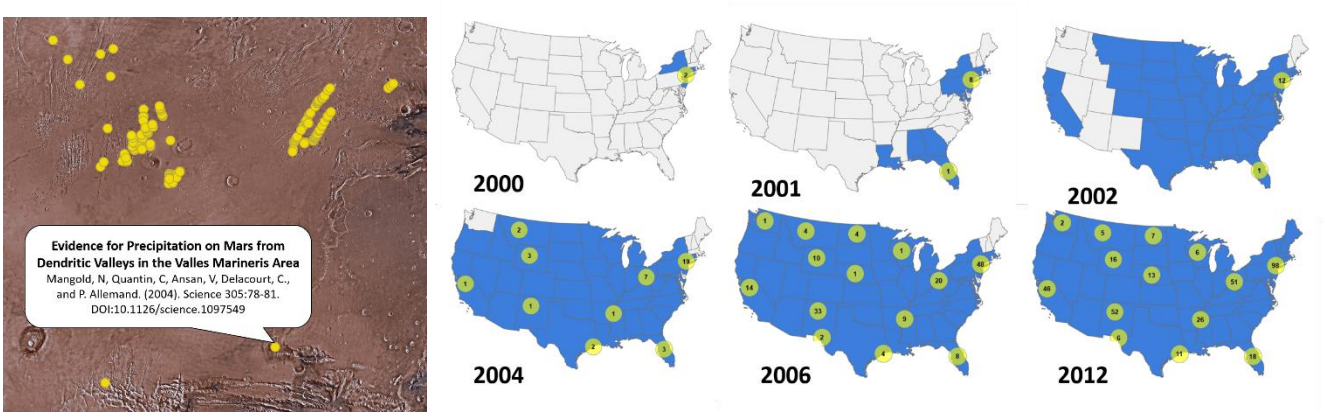

Figure 3. The concepts developed through this proposal are extensible beyond agricultural and environmental literature to any knowledge domain where location is important including planetary research (Mars example, left) and infectious disease studies (Right, tracking progression of West Nile Virus across the US [blue shading] and published studies on the disease's impacts [yellow circles]).

The deliverables from this phase, like the application to result from the successive phase, will be made publicly available. The governing philosophy of our team is one of open source development, even though we are exploring ESRI technology for its role in our designs. We plan to make the designs public, the system adhering to W3C standards for web APIs, and provide bibliographic metadata in standard serializations such as JSON-LD and XML, as well as content structures like JATS, or BibTeX.

Our sustainability goals for the project are to achieve at minimum the necessary design information required to pursue future work. While we are planning to pursue an NLG project grant, we anticipate continuing to work on the project with interested stakeholders regardless. Once the planning grant is complete, we will have the necessary specifications to pursue funding from numerous sources to try to continue the work if a project grant is unsuccessful. In both the case of the CRDB and JournalMap, the current iterations, while not ideal, are sustainable and have been for some time. We anticipate that they could remain in their current states indefinitely, although that would miss an opportunity.

Jesse Shera (1961) claimed that the purpose of librarianship is to "maximize the social utility of the graphic record". Ranganathan (1931) preceded him by imploring those in libraries to "save the time of the reader." We strongly believe the deliverables of this project – automating geotagging and metadata assignment, providing context-similar searching, an open API to re-use the tagged literature, models for integrating with other bibliographic resources – can produce a positive economy of scale in terms of cost and efficiency for librarians and of access, precision, and recall for researchers that use it. Libraries are the information organizations that often seem to be heavily constrained by institutional budgets and bureaucracy that slow systemic innovation, reduce risk-taking and experimentation, which hamper efforts to change the way their users access information. We feel that this project has an ability to contribute to the library community in a manner that lowers costs, enhances access to knowledge, improves discoverability of spatially-oriented literature, data, and other digital resources in ways that follow Shera and Ranganathan's direction: maximizing the utility of the agricultural and environmental record and saving the time of everyone involved.

Task Name	Sept '20	Oct	Nov	Dec	Jan '21	Feb	Mar	Apr	May	Jun	Jul	Aug '21
Goal 1: Create Application Architecture and Prototypes												
Objective 1: Reimagine existing architecture for JournalMap and CRDB												
Objective 2: Test Improved Approaches to Geotagging												
Objective 3: Develop CRDB prototype with JournalMap												
Goal 2: Enlist Stakeholders for a Project Grant												
Objective 4: Engage Stakeholders in Design and Prototyping												
Objective 5: Conduct Needs Assessment and Requirements for Stakeholders' Integrations												
Goal 3 – National Leadership Project Grant Proposal												
Objective 6: Project Grant Proposal												

DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

INSTRUCTIONS

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE

Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA

Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). Explain and justify any prohibitive terms or conditions of use or access, and detail how you will notify potential users about relevant terms and conditions.

We will make all data, algorithms, and code open and publicly accessible. We will assign a Creative Commons CC BY 4.0 license to all outputs. Everyone involved in this project is committed to an open and accessible set of resources to facilitate interest and re-use of this project's outcomes. The project website, hosted at UI's Northwest Knowledge Network and github repositories will be accessible to the public.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

We will make all data, algorithms, and code open and publicly accessible. We will assign a Creative Commons CC BY 4.0 license to all outputs. No further conditions will be imposed on access or use. Any publications resulting from this work will be published in open access outlets, or a copy will be made available through the University of Idaho or Kansas State University institutional repositories.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

N/A

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

For Objective 1 - we will create system architecture diagrams, system requirements documents, and diagrams of database schemas. For each of these, we will use diagramming and modeling software, such as Adobe Creative Cloud (Indesign, XD). They will be minimal in file size and we will create final PDF/A versions of the files aside from the internal, proprietary formats.

A.2 List the equipment, software, and supplies that you will use to create the digital content, resources, or assets, or the name of the service provider that will perform the work.

We will use resources and staff at the Northwest Knowledge Network, a data services service center within the University of Idaho. Their staff has access to all of the tools required for diagramming and wireframing - e.g. Adobe Creative Cloud - and they maintain a computing infrastructure including development and production virtual machines and servers, live web hosting, application development support, and full server administration. Code will be hosted within a University of Idaho Github repository and architechture disseminated through the project website.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

Architecture/diagram files will be in PDF/A formats.

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

Diagramming and system architecture will occur iteratively by experienced personnel who have a proven track record of creating and designing systems. Feedback and quality control will also be provided by the stakeholder advisory group, who can provide direction on the appropriateness of the approach from the perspective of librarian and patron experiences.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

All resources will used the aforementioned NKN systems to manage files and folders. [something about NKN's backup]. As this project is intended to support a future project, it is anticipated that the tools will persist in their location for the foreseeable future. Current/outdated iterations of both JournalMap and CRDB are maintained currently on Amazon Web Services' and KSU's servers respectively with funding provided by the project participants. We expect that to be a fallback option should we be unable to maintain the resources on NKN. Any developed code for new versions of JournalMap and the CRDB, from which new iterations could be deployed, will be maintained on a public code repository (e.g., GitHub).

Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata or linked data. Specify which standards or data models you will use for the metadata structure (e.g., RDF, BIBFRAME, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

Content metadata will not be necessary for the diagrams and schemas, as they are described by the context of the web sites or reports in which they exist.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

n/a

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

Since this planning grant is primarily about design, we will not actively pursue metadata sharing strategies. Our hope is for a future project in which we can address improving access to the JournalMap index. We will work with stakeholders to build relationships that will use our resources and engage in direct outreach through conference attendance and contacting prospective partners.

Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content, delivery enabled by IIIF specifications).

All project documents will be openly available online through the project's websites. The code and scripts may start in a private repository during the period of active development, but will be made public by the end of the period of performance.

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

JournalMap: https://www.journalmap.org
Croplands Research Database: https://www.lib.k-state.edu/gracroplands
JournalMap Geoparsers and other code: https://github.com/JournalMap

SECTION III: SOFTWARE

General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

Objective 2 - 1) an algorithm to georeference study locations within journal articles, and 2) an algorithm to conduct spatial searching within a web interface. The first will permit automated parsing of journal text, identification of study area coordinates or place names, and pass those coordinates or names geocoding service. The features extracted from the data layers will then be associated with the article in the database for retrieval in searching and browsing. The second algorithm will test ESRI ArcGIS tools for performing spatial searches and feature extraction. The audience is other developers. Objective 3 - we will create a prototype web interface to demonstrate integration of the CRDB and JournalMap. This will demonstrate how the CRDB corpus can utilize JournalMap's features, while providing users to search articles, filter by facets, and follow links to articles. The audience is anyone seeking to search for environmental literature.

A.2 List other existing software that wholly or partially performs the same or similar functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

ScienceBase, Pangaea, BioStor, and DataOne all provide similar spatial searching. Only Biostor and the GLOBE project appear to do what JournalMap does with automated parsing. We feel that existing software does not fully leverage the opportunities in natural language processing that has been advanced over the past decade. Our software would integrate these advances and improve the scale of article indexing in terms of both magnitude of articles and frequency.

Technical Information

B.1 List the programming languages, platforms, frameworks, software, or other applications you will use to create your software and explain why you chose them.

The current web interface of JournalMap is in Ruby on Rails with the data stored in MySQL relational database tables. Most of the data handling, feature extraction, geoparsing is written in Python. The search index for JournalMap is ElasticSearch. The CRDB is a web application served through PHP. The new prototype JournalMap/CRDB system will be developed in Drupal and PHP, and the data stored in MySQL tables.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

This project is designed to test out possible strategies and propose an application architecture. Any software is not intended to interoperate with other systems at this time.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

All Python libraries used for testing will be articulated in an accompanying requirements.txt document, such as the *re* library for regular expression matching or *arcpy* for use of ESRI ArcGIS Python tools.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

Software developed in this project is entirely developmental. It is meant to experiment on ways of geotagging articles and to demonstrate possible approaches to a search interface. Therefore, the software will not be considered "production-level" or a product for formal distribution. Code contributors will be expected to comment their contributions, and administrators will ensure that appropriate documentation is generated. Once code is considered complete, it will be linted according to the appropriate standards, such as PEP8 for Python code and JES6 for Javascript.

B.5 Provide the name(s), URL(s), and/or code repository locations for examples of any previous software your organization has created.

JournalMap: https://github.com/JournalMap
Northwest Knowledge Network: https://github.com/northwest-knowledge-network
UI Libraries: https://github.com/uidaholib
KSU Libraries: https://github.com/kstatelibraries

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

All code will be made available through JournalMap's Github repositories.

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

JournalMap on Github.com

URL:

https://github.com/JournalMap

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

A.1 Identify the type(s) of data you plan to collect or generate, and the purpose or intended use(s) to which you expect them to be put. Describe the method(s) you will use, the proposed scope and scale, and the approximate dates or intervals at which you will collect or generate data.

For Objective 2: a bibliographic dataset of articles listed in the CRDB. Code and scripts for performing the extraction and geotagging written in Python 3+. These scripts will produce approximately 6 datasets: 1) benchmark, manually geotagged, 2) LocateXT results, 3) current parser + LocateXT results, 4) current parser + spaCy, 5) current parser + NLTK, 6) current parser + Stanza. There will also be summary statistics produced as CSVs for visualization and analysis of error rates, as well as graphs and charts. All are expected to be produced using scientific Python libraries and Jupyter notebooks. The timeframe for this is between Sept 1, 2020 - Feb 1, 2021.
For Objective 5: survey results of users of the CRDB to inform user interface prototyping. Analysis of the survey results will be in a Word document with the possibility of Excel graphs and a CSV of summary data. The timeframe will be Feb 1 - June 1, 2021.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

While our assertion is that the survey does not ask about human subjects information, we will submit a proposal to the University of Idaho IRB to verify that it is in fact exempt. We will seek to make our survey results public, for full transparency, but will ultimately respect the decision of the IRB committee.

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

We will not ask for sensitive or personally-identifiable information.

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

Replicating the geotagging test results will require use of Python version 3.x+. Depending on the method used, it will require the appropriate (open source) libraries spaCy, NLTK, or Stanza.

However, viewing and retrieving the results or code will require no special software. File outputs will be in either CSV or JSON format, widely accessible by any text editor. Code will be in Python.

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

Procedural steps will be established for testing the parsers and documented in metadata accompanying the results. Code will be commented with necessary information to understand the scripts. The survey results will have a DDI Codebook XML file associated with it that documents the method, instrument, general information, survey assumptions, and responses.

All data and metadata will be placed together in the public data repository managed by the Northwest Knowledge Network (the institutional data repository for the University of Idaho). It will be accessible to all project participants in perpetuity.

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

All data and metadata will be placed together in the public data repository managed by the Northwest Knowledge Network (the institutional data repository for the University of Idaho). It will be accessible to all project participants in perpetuity.

A.7 Identify where you will deposit the data:

Name of repository:

Northwest Knowledge Network (University of Idaho)

URL:

https://data.nkn.uidaho.edu/

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

The DMP will be reviewed by the Project Director or designee in Feb 2021 to ensure we are complying with our assertions. If not, we will address the deficiencies. If new technology or resources are used, we will add them to the plan accordingly.