Request for Proposals

Software Engineering Services for Web Monitoring System for Invasive Species

June 21, 2016

Overview

The Great Lakes Commission (GLC) invites proposals from qualified applicants to provide software engineering services to develop updates and enhancements to an existing web-crawling software system. The existing system – Great Lakes Detector of Invasive Aquatics in Trade (GLDIATR) – searches the web to identify websites offering invasive species for sale. These services will be provided as part of an overall project for which the GLC has received funding from the U.S. Environmental Protection Agency's Great Lakes Restoration Initiative program (glri.us). The requested services are expected to be completed by January 31, 2017 with maintenance as needed through January 31, 2018.

Funding

Up to $100,000 is available for the requested services.

RFP Process and Deadline

Proposals must be submitted via email in PDF format to Erika Jensen (ejensen@glc.org) at the GLC by **Friday, July 22, 2016, at 1:00PM Eastern (12:00PM Central)**.

- Pre-submittal conference call: Tuesday, June 28 2016
- Proposals due: Friday July 22, 2016 (by 1:00pm Eastern)
- Interview finalists: July 27 – August 3, 2016
- Final selection: Friday, August 5, 2016

A one-hour conference call will be held on **Tuesday, June 28, at 9:00AM Eastern (8:00AM Central)** to answer questions about the proposed work. Interested parties should contact **Erika Jensen at the GLC (734-971-9135, ejensen@glc.org) for more information, including instructions on how to participate in this call.**

Following the submittal and review of proposals, the GLC will select up to three finalists. The GLC will conduct interview(s) with these finalist(s) at a mutually agreed upon time during the period of July 27 – August 3. Following the interview(s), the GLC will notify finalists of the decision.

Who Should Apply

Independent consultants, firms, universities and other interested parties (or consortiums thereof) with documented experience in computer software engineering, web technology including data mining, text analysis, and other relevant fields. Web data extraction experience is also a plus. In compliance with the U.S. EPA Disadvantaged Business Enterprises (DBE) program, all DBEs, including Minority Business Enterprises (MBE), Women's Business Enterprises (WBE) and others, are encouraged to apply.

The GLC, as an equal opportunity employer and recipient of federal funding, complies with applicable federal and state laws prohibiting discrimination. It is the policy of the Great Lakes Commission that no person employed by or doing business with the GLC shall be discriminated against, as an employee or applicant for employment, because of race, color, national origin, religion, age, sex, height, weight, sexual orientation, marital status, partisan considerations or a disability or genetic information that is unrelated to the person's ability to perform the duties of a particular job or position.

Contact

Erika Jensen, Project Manager, Great Lakes Commission
2805 S. Industrial Hwy., Suite 100, Ann Arbor, Michigan 48104
Phone: 734-971-9135 x139, Email: ejensen@glc.org

Background

The Great Lakes Commission (GLC) is a public agency, founded in state and federal law, and dedicated to coordinating and fostering a regional approach to Great Lakes management. The provinces of Ontario and Québec participate as associate members. The GLC assists its member jurisdictions in speaking with a unified voice and helps them collectively fulfill the vision for a healthy, vibrant Great Lakes-St. Lawrence River region. Functions of the GLC include communication and education, information integration and reporting, facilitation and consensus building, and policy, coordination and advocacy. The GLC is governed by delegations from each of the Great Lakes states and from Ontario and Québec. The GLC's professional staff, based in Ann Arbor, Mich., provides technical, communications and managerial support for diverse environmental and economic initiatives. See www.glc.org for more information.

The GLC is leading this project to support aquatic invasive species (AIS) prevention efforts by addressing the sale of invasive species over the Internet. In 2012, GLC contracted RightBrain Networks (www.rightbrainnetworks.com) to develop the GLDIATR system to assess the availability of invasive species via Internet sales, identify sellers, and support the activities of AIS management agencies. The GLC received a second grant from U.S. EPA to continue this work. Previous funding allowed for the development and testing of GLDIATR, serving as a "proof of concept" for the application of this technology to assist with invasive species prevention. The GLC will now build on this progress to advance solutions to the Internet trade of AIS through further development of GLDIATR.

The GLDIATR system automatically searches the Internet and collects the URLs of web pages offering invasive species for sale. GLDIATR is operated on hardware purchased and housed within the GLC and is running KVM-based virtualization against CentOS. GLDIATR code is written using Python (2.7.X), Javascript and HTML code. Other programming elements of GLDIATR include MySQL and Mongo databases, Natural Language Tool Kit (NLTK) for language processing, RabbitMQ for message queue, Celery for task queue, and Pyramid web framework with Mako template library and Bootstrap frontend framework.

The GLDIATR system is comprised of several components (or applications). The front-end components include a web-based dashboard user interface that allows an administrative user to schedule web-crawling, conduct classification model training, manage users and data, and view data reports, among other tasks. Back-end applications within the system carryout the following tasks in a distributed environment:

- **Datastore application:** Responsible for data retrieval and storage. Other applications can make requests to save information and to view existing information.
- **Retrieval application:** Responsible for performing Internet searches and downloading web pages. Other applications can make requests for specific searches to be performed and for specific URLs to be downloaded.
- **Preprocessing application:** Responsible for performing species matching on pages. The application uses natural language processing to break apart webpage components and remove unwanted content in order for other applications to target the most relevant sections. Other applications can request that specific downloaded pages be processed.
- **Learning application:** Responsible for turning processed pages into data structures that can later be classified, and to build data structures that inform algorithms and the classifier application. Other applications can make requests for pages to be turned into data structures and to send lists of pages to use as data on which to build the classifier data structure.

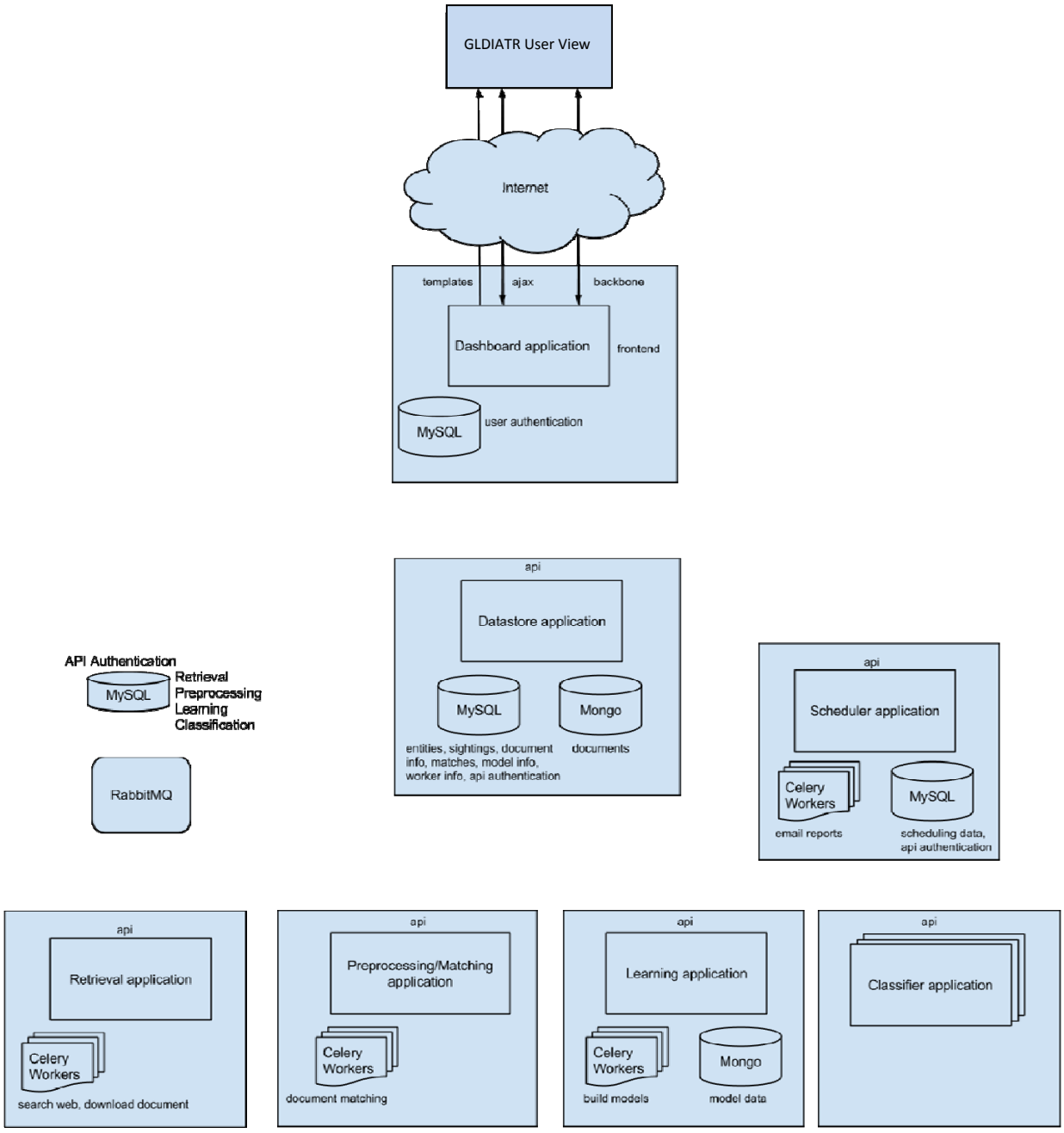
- **Classifier application:** Responsible for running the classification data structure against page data structures produced by the Learning application. Other applications can request specific downloaded pages be used to build the classifier data structures and algorithms that perform the classification. It also accepts requests for specific pages to be classified using the current classifier data structure.
- **Scheduler application:** Responsible for running searches and producing reports at specified times. Other applications can request specified tasks repeat and run at specific times.

These applications utilize communication protocols that are designed and implemented through APIs and message queues. The GLDIATR work flow through theses various applications typically proceeds as follows:

1. Search: Perform a search in order to get a list of URLs and, for each URL, repeat steps two through five
2. Retrieve: Download the page
3. Preprocess: Inspect the page text for search terms and, if one or more are found, continue with steps four and five
4. Learn: Break the text of the page into a data structure
5. Classify: Run the classifier application on the data structure to determine the page is a positive (offering a sale) or a negative (no sale offered) hit and save the result

GLDIATR performs its searches of the Internet through the APIs provided by the search engine Bing, the major online retailers Amazon and eBay. All searches are executed based on a query (i.e., set of search terms) set within the GLDIATR Dashboard application. GLDIATR queries simulate an individual independently accessing Bing.com, Amazon.com or eBay.com and typing a query into the search box. For purposes of this project, the search terms are defined through a “species watch list” – developed by the GLC – of common and scientific names of potentially invasive species.

Figure 1. GLDIATR System Architecture



Scope of Services

The existing GLDIATR software was developed to demonstrate the application of web-crawling and web data mining technology to the problem of invasive species trade over the Internet. The intent was to help bring clarity to the issue and illustrate opportunities for solutions. The selected contractor will assist the GLC by developing new, added features to enhance the GLDIATR system. These enhancements include:

- Geographic Location Algorithm: During the previous project phase, GLC staff manually collected information on the geographic (physical) location of online retailers, based on the results returned by GLDIATR. The GLC requests that the selected contractor develop a new algorithm for GLDIATR that would automate this process and provide this additional information to users. The GLC anticipates that development of this algorithm would be straightforward for marketplace websites such as eBay and Amazon in which location information is placed on the page in a consistent manner that is relatively easy to predict and identify. It is likely that significantly more development and testing will be needed to automatically extract this information from various other online stores in which the placement of geographic indicators (e.g., contact information) varies from site to site. Prospective contractors should include in their proposal how they will work with GLC staff to set and achieve goals for this algorithm, recognizing the inherent challenges of this task.

- User Interface: In order to facilitate and streamline user engagement with GLDIATR, further customization of the user interface is needed. Currently, GLDIATR reports the entirety of results from system searches in two "reports:" a species report including sale pages sorted by species, and a domain report including sale pages sorted by primary domain (e.g., www.ebay.com). Both of these reports allow the user to filter the results by a specific subset of species, species regulatory status, date the page was found, and other parameters. The user manually sets the filters based on their specific interests. The default view shows all the pages identified over the life of system operation. As envisioned, the new user interface would allow users to establish default filters based on their locations and species of interest (e.g., jurisdictionally regulated species). These filters would be designed to allow users to select a subset of species from the existing watch list, by individual species or by other characteristics such as federal and state regulatory status or taxonomic group (e.g., fish, plant or invertebrate). Further, users would also be able to select to only see sale pages from sellers found to be located in one or more countries (e.g., U.S., Canada, Australia), states (e.g., Illinois, Indiana, Michigan), or provinces (e.g., Ontario, Quebec). The default view would always be to see everything (i.e., all species, all sellers).

  The GLC would work with the contractor and users to identify specifications for the unique interfaces. These interfaces would also be developed so that they can be modified in the future to accommodate evolving species regulations or interests. These updates should not change the base functionality of GLDIATR which will continue to search the Internet for all species of interest and sellers; it will merely modify what certain individual users see in their reports. The ability of a user to view all results (not just what is included in their customize view) would be retained.

- Case Management: GLDIATR was conceived, developed and designed to help facilitate and target activities that will block the organisms in trade pathway from leading to AIS introductions in the Great Lakes region. Tracking the implementation of these activities is an important part of measuring progress and evaluating success. To facilitate this, the GLC is proposing to add components to the GLDIATR software that will allow any user to record their activities. This is likely to include flagging sellers that have been contacted regarding species in their inventory identified by GLDIATR, either in an enforcement or general education-outreach capacity, including recording the date when the notification was sent, how the notification was sent (e.g., email, contact form, phone, U.S. mail, etc.), and whether a response was received. It will also be important that users have the ability to share this

information with each other to ensure management is coordinated and duplication of effort is reduced. The format and specific features will be developed and tested with the GLDIATR Advisory Board. The GLC strongly recommends that contractors consider the use of existing, popular third-party case management platforms for this task. The cost of any proposed commercial software, including acquisition and maintenance through the end of 2018, should be included in the proposal.

The selected contractor will be expected to carry out this work under the direction of and in consultation with the GLC. This consultation will require regular conference calls and/or web conferences, with the potential for in-person meetings, if needed. The selected contractor will be required to provide complete documentation and installation documentation. If non-standard libraries or other technology are used they will need to be included in the installation package and be freely transferrable to other parties.

To facilitate the development of proposals, the GLC will provide access to GLDIATR source code and documentation to applicants upon request and subject to a non-disclosure and confidentiality agreement.

Schedule
The selected contractor will be expected to carry out the requested services over a period of approximately eighteen (18) months (August 2016 – January 31, 2018), with primary development work being completed within the first six months (by January 31, 2017).

Requirements for Proposals
Proposals must be no longer than twelve (12) pages. Appendices totaling no more than 10 additional pages are allowed and should include resumes or CVs for personnel that will carry out the work. Proposals must include the following information:

1. Name of applicant organization and principal investigator, including contact information
2. Summary of applicant's qualifications including experience in computer software engineering, web technology, web data mining, text analysis, machine learning and other relevant fields
3. Concise examples of past work demonstrating knowledge of and experience working in the fields mentioned above
4. A statement demonstrating the applicant's understanding of the services requested
5. A description of how the outcomes described in the scope of services will be achieved, including proposed hardware needs, software to be used, a schedule and budget.

The proposed budget should include salaries, fringe benefits, other direct costs needed (including travel) and indirect costs.