Developing an open access croplands research database through global collaboration

Livia Olsen, Tara Baillargeon, and Harish Maringanti

Published Version Information

Citation: Olsen, L., Baillargeon, T., & Maringanti, H. (2012). Developing an open access croplands research database through global collaboration. Journal of Agricultural & Food Information, 13(1), 35-44.

Digital Object Identifier (DOI): doi: 10.1080/10496505.2012.639272

Publisher's Link: http://www.tandfonline.com/toc/wafi20/13/1

Developing an Open Access Croplands Research Database through Global Collaboration

LIVIA OLSEN, TARA BAILLARGEON, and HARISH MARINGANTI

Kansas State University Libraries, Kansas State University, Manhattan, Kansas, USA

This paper describes the processes, challenges and outcomes of a project undertaken by Kansas State University (K-State) Libraries and a global community of researchers. This project, initiated by librarians in the newly created Faculty and Graduate Services Department, involved collaboration with a K-State agronomist. The initial concept was to create an open-access database of croplands research submitted by researchers from the Global Research Alliance Croplands Research Group, a consortium of over 30 countries. Due to the project's complexity, it was determined that a manageable approach was to pilot the project by only including research from the United States and Australia.

KEYTERMS croplands, databases, collaboration, solr, Global Research Alliance on Agricultural Greenhouse Gasses, open access, scholarly communication, library collaboration, faculty collaboration

---

Received 10 October 2011; accepted 21 October 2011.

The authors extend their thanks to Tara Mosier for the graphic design she created for the website and Dhanushka Samarakoon for his programming assistance.

Address correspondence to Livia Olsen, K-State Libraries, 137 Hale Library, 1100 Mid-Campus Drive, Manhattan, KS 66506, USA. Email: livia@k-state.edu

INTRODUCTION

In 2009, K-State Libraries embarked on reorganization in response to the changing roles of librarians. One outcome of the reorganization was the creation of a Faculty and Graduate

Services Department whose mission, in part, is to form collaborative partnerships with faculty to enhance research on campus. One of the first projects initiated by the new department involved collaborating with Dr. Chuck Rice, a K-State Agronomist. This partnership expanded to include several members of the Libraries' faculty and staff, as well as researchers from the United States and Australia to create an open access Croplands Research Database pilot.

BACKGROUND

To get a better sense of the types of research activities occurring on campus and to identify potential roles for the library, Faculty and Graduate Services librarians networked by attending campus research meetings. At one such meeting, Faculty and Graduate Services librarians met Dr. Chuck Rice, a K-State agronomist. At this meeting, discussions began with Dr. Rice about how the Libraries could collaborate with him and members of the Global Research Alliance (GRA) on Agricultural Greenhouse Gasses. Dr. Rice is a U.S. representative on the GRA, a global alliance of more than 30 countries established in 2009 as a result of the United Nations Climate Change Conference in Copenhagen, Denmark. One of the key issues addressed by the GRA is the need to "to limit the loss of carbon and nitrogen from crops and soils to the atmosphere" (Global Research Alliance on Agricultural Greenhouse Gases, n.d., ¶ 1). The GRA consists of three research groups: paddy rice, livestock, and croplands. Dr. Rice is affiliated with the croplands group and recognized the importance of making this research freely accessible to support the work of researchers on a global scale. Dr. Rice approached the Libraries about joining him in collaborating with his international colleagues in the GRA to develop an open access database of research about greenhouse gas emissions from croplands.

Librarians collaborating with a faculty member to create an open access database of croplands research aligned with the goals of the Faculty and Graduate Services Department, making participation by the Libraries desirable. K-State Libraries' Administrative Team agreed that the croplands database was a project that would benefit from librarian expertise in collecting and organizing information, as well as from the strong technical capabilities of the Libraries' staff. Before embarking on the project, the team needed to ensure that the proposed database would not result in a resource that was otherwise available. The last thing the team wanted was to replicate something that already existed. However, it soon became apparent that they would be creating a unique resource, because the proposed croplands database included a variety of crop-related resources not otherwise available.

VISION

Dr. Rice came to the project team with a distinct vision from the GRA Croplands Research Group. First, he envisioned creating a publicly available database of the literature about agricultural greenhouse gasses in croplands. Second, he was interested in creating an additional database containing a list of all research projects, grants, and researchers in the GRA member countries. Third, he wanted to create a repository of cropland studies data. He realized that achieving all three of these goals simultaneously would be a monumental task, so he asked the team to focus first on creating the literature database.

The goal of the literature database is to compile information about croplands and greenhouse gases from all over the world. The database will contain literature about greenhouse gases and many different crops located in various climates. In addition to journal articles, the content will include conference papers, white papers, videos, data sets, and any other useful

research files. The database will also serve as a place to preserve grey literature that might be difficult to find on the Internet or that might disappear over time. Another benefit of the database is that it will be open access. Users, regardless of their affiliations, will be able to access the database~~'s~~ content. Information will not be hid in a fee-based database that those at under-financed institutions may have difficulty accessing. When applicable, the team will work with publishers and authors to ensure copyright compliance. The open access nature of the database, as well as the potential for numerous collaborators to contribute content, presents some challenges. The project team will need to ensure that there are no copyright violations when citations, abstracts, and/or full-text articles are submitted by researchers who may not understand copyright restrictions.

PILOT PROJECT

The team determined that the best approach to creating the database was to start small. Dr. Rice was working closely with a GRA Croplands colleague in Australia, so the project team decided to start the pilot project with literature from the United States and Australia. This ensured the opportunity to find and correct any problems with the pilot project before opening the site up to accepting literature citations from the rest of the GRA countries. Starting with a small amount of data gave plenty of time to understand challenges involved in editing the data and identify efficient solutions. The pilot project also gave the team time to identify and correct any copyright issues and set up an efficient workflow procedure for international colleagues to follow.

The project team began, on paper, to design what they felt were the necessary components of the database and its accompanying website to display the content in the database. The most important features identified by the GRA members were the abilities to search and

browse the database by country, climate, and cropping system. A preliminary metadata scheme was formulated based on the data and initial requirements. Features of the public website, content submission form, and staff interface were selected for simplicity and user friendliness, such as adding mouse-over, context-sensitive help menus and listing subject headings in drop-down menus to reduce errors. The most challenging part was developing the controlled vocabulary for country, climate, and cropping system. The climate terminology was the most difficult. The team utilized the Köppen Climate Classification System, which is over 100 years old and developed in many different iterations requiring extensive sorting to develop a controlled vocabulary. A 2007 article by Peel, Finlayson, and McMahon helped the team to understand this climate classification system and shaped the terminology used in the controlled vocabulary for climate.

Fortunately, K-State Libraries has a graphic designer on staff that created a simple, eye-appealing design for the public website. The project team communicated with Dr. Rice and his USDA colleagues to elicit feedback about the design. Before viewing the design they had no opinion about what the site would look like. However, once they saw the design (Figure 1), the Libraries' project team received suggestions from them that were very helpful.

IMPLEMENTATION

Manageable technical goals were defined for the pilot project to test the feasibility of the team's approach and to evaluate the selected technical platforms. Feedback from the shareholders was important in designing the system and it was decided, early on, to focus mainly on the user interface design for this initial phase. Getting data, in the form of citations, abstracts, full text, and data sets, into the system was equally important, and it was essential to create tools for the

staff members to ingest and edit data (see Figure 2). Thus the goals for the pilot project revolved around (i) creating an easy-to-use interface for the users and (ii) creating tools for staff workflows.

The project team envisioned that the sources of data for this project would be varied. For the pilot, the team decided to use a single data set to experiment with during the pilot phase. Major decisions on the back-end database structure were deferred to the subsequent phases. Iterative software development process fits well with this approach as, after each phase and more data, the system can be modified appropriately based on the feedback.

With Web 2.0 and minimalistic design gaining traction, the project team wanted the main public interface to be as free of clutter as possible. Keeping this in mind, the search engine was made the focal feature of the public interface. All the features that a modern search engine can support, such as sorting, filtering, spell checking, and full-text search, will be implemented, but, for the pilot phase, the basic requirement was to setup a simple search feature that returned and ranked results quickly. To help navigate the search results, the team was interested in exploring the faceting option as this feature would allow users to narrow down results in an appropriate manner. Apache Solr (2011) was selected as the site's search engine, as it seemed to fit the bill perfectly and the project team could also rely on its prior experience in using Solr in other projects. The technology stack used to develop the pilot project included Apache, PHP, MySQL (2011) and Solr.

WORKFLOW

On the staff workflow side, emphasis was on the data ingest. A simple, easy-to-use interface was critical for the staff members to enter the data. A bulk import process that could ingest multiple

records was desirable, as it would help in optimizing resources. This would enable librarians to focus on describing and mapping the content, rather than focusing on data entry operations. To limit human errors and promote consistency, fields were pre-populated wherever appropriate. The pre-populated fields were displayed using drop-downs and check boxes. As this project focused on the creation of a specialized database dealing with agricultural greenhouse gases in croplands literature, the usage of controlled vocabulary was important to describe and classify climate types and cropping systems.

For the pilot project, Dr. Rice's personal literature database was utilized. The initial data set, in the form of an EndNote file, was ingested into the system using the bulk import process. To create mappings between the database fields and the input data, classification from librarians was necessary. Once the data was ingested, librarians edited the ingested data and added subject headings to reflect the country, climate types, cropping systems, and keywords.

Close to 1,200 records were processed as part of the pilot project. The system has basic search features and faceting options as previously discussed. Limiting the search scope in the pilot phase means that full-text searching is not yet available. In the current setup, the user has the option of narrowing the results by choosing faceting options among cropping systems, climate types, country and keywords.

It was quickly discovered that some of the data had not transferred correctly and needed extensive cleanup. This required an hour to clean up approximately 12 records, a very slow process which highlighted the need for more help on the project. The team is working to correct any problems with mapping the data from its original format into the database and is pursuing hiring a student worker to help with the project. Dr. Rice's contact in Australia recently sent a

list of Australian literature on croplands and greenhouse gases, so international content will soon be added to the database.

Adding a student to the project team will give the team an additional viewpoint on the project. This will help the team to evaluate the feasibility of scaling-up the workflow when many contributors are entering information into the database. Dedicating someone to work on cleaning up records on a daily basis will help uncover problems with the workflow that may need to be corrected. A workflow tracking feature is a must for both student workers and international contributors. Given the potential number of countries and people who will be involved with the project, the ability to monitor the status of a record is essential. The workflow tracking feature will be added in subsequent phases of the project.

NEXT STEPS

The pilot of the GRA Croplands Database was created using literature from the United States and Australia, and the next step is to assess the database by seeking feedback from GRA members. Modifications will be made to the database based on this feedback. Though an assessment has not yet been conducted, the project team already has plans for improvements to both the database and the website. Streamlining the data entry process will be a primary focus.

Efficient processing tools for staff are needed to improve the quality of the data. Basic editing capabilities were addressed in the pilot project, and the team plans to add some advanced options in the later phases, including the option to bulk-edit records. In the pilot phase, interfaces allowing the public and staff to retrieve and edit information were the focus. Once other member countries begin to contribute data, the database structure will be finalized and support for a standard metadata schema for ingesting data will be developed. After full-text documents are

submitted to the database, the project team will enhance the current search feature to include full-text indexing. The team will also explore options to make this data available to other systems using open protocols such as Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). (Open Archives Initiative, n.d.)

To scale-up the project to include research from all GRA member countries, the project team will seek grant funding to support further development. The ongoing maintenance related to the project will be very time-consuming, so funding to support positions that will work on the project is crucial. The project team intends to develop a promotional plan that encourages member countries to supply the database with more data, while also encouraging researchers to utilize the information in the database.

CONCLUSION

Throughout the course of this project, many lessons were learned, the first being the importance of librarians getting out of their offices and networking with campus researchers. By attending open research meetings, departmental brown-bag sessions, and other events on campus, librarians were able to stay informed about current and emerging research projects. Their presence at these events encouraged them to think about ways in which the library could contribute to research on campus beyond the service of traditional research consultations.

The second lesson was to start with a manageably-sized project, even when potential collaborators set very ambitious goals. Best practices dictate leveraging existing resources and skills available in the library. The project team was able to use existing EndNote files of citations from researchers and expertise gained using the Solr search interface on a previous project.

The third lesson was that buy-in from the library's administration is essential. Administration has the ability to allocate time and resources to the project. This became a truly interdepartmental project, further highlighting the importance of having the administrative support that allowed members from multiple departments to contribute. The GRA Croplands Research Database is now available at http://www.lib.k-state.edu/gracroplands.

REFERENCES

Apache HTTP Server (Version 2.2.15) [Computer software]. (2011). Forest Hill, MD: Apache Software Foundation.

Apache Solr (Version 3.1.0) [Computer software]. (2011). Forest Hill, MD: Apache Software Foundation.

Global Research Alliance on Agricultural Greenhouse Gases. (n.d.). *Croplands research group*. Retrieved July 25, 2011, from http://www.globalresearchalliance.org/research/croplands-research-group/

MySQL Community Edition (Version 5.1.45) [Computer software]. (2011). Redwood Shores, CA: Oracle Corporation.

Open Archives Initiative. (n.d.). *Open archives initiative protocol for metadata harvesting.* Retrieved October 26, 2011, from http://www.openarchives.org/pmh/

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions*, *4*(2), 439-473.

PHP (Version 5.2.13) [Computer software]. (2011). The PHP Group.

FIGURE 1 A screen shot of the GRA Croplands Research Database homepage with a simple search box in the upper right.

FIGURE 2 A screen shot of the GRA Croplands Research Database data input form which allows for edits submitted through a bulk import process.