This is a preprint of an article whose final and definitive form has been published in Library Hi Tech News [30(4):1-5, 2013]; Works produced by employees of the US Government as part of their official duties are not copyrighted within the USA. The content of this document is not copyrighted. Library Hi Tech News is available online at: http://dx.doi.org/10.1108/LHTN-03-2013-0017

Crowdsourcing: Divide the Work and Share the Success

Constance J. Britton, Ohio State University, OARDC Library, Wooster, OH
Allison V. Level, Morgan Library, Colorado State University, Ft. Collins, CO
Melanie A. Gardner, National Agricultural Library, Beltsville, MD

Introduction

In the library world, everyone is busy with more and more work-related projects. Despite our best intentions, sometimes even valuable projects get set aside. In an effort to move a project towards completion, a novel organizational approach of "crowdsourcing" was tried.

The Agriculture Network Information Center (AgNIC) is, "a voluntary alliance of members based on the concept of 'centers of excellence'. The member institutions are dedicated to enhancing collective information and services among the members and their partners for all those seeking agricultural information over the Internet. Joining forces to enhance impact and deliver increasing access to information and expertise, it enables partner institutions to make the most of available resources and increase impact." (Agricultural Network Information Center , 2012)

A primary activity by AgNIC partners is to identify web-based agriculture and natural resources information and contribute associated metadata which is aggregated into the AgNIC portal. On the portal, keyword terms entered in the AgNIC search box retrieve metadata from the AgNIC resource database, which indexes AgNIC partner created websites, other web content, digital objects harvested from institutional repositories (IRs), news, calendar events, and more. In order to further the AgNIC mission, partners also serve on committees organized around key projects. The AgNIC Content Committee is charged with identifying additional content that is in scope for the portal. During the 2011 AgNIC annual meeting, the committee sought to identify an imaginative way to spur participation and create a review process to identify more relevant web content. The committee chair proposed a crowdsourcing project to accomplish this task.

Crowdsourcing is a distributed problem-solving technique leveraging the efforts of a group, known as "the crowd". A project is defined and volunteers are invited to contribute to its accomplishment. The volunteers are dispersed and may not even be members of the organization. Comparing the definition of crowdsourcing with the AgNIC mission proved a logical mash-up.

The AgNIC Mission as Crowdsourcing		
Definition: Crowdsourcing is a *distributed problem solving and production process* that involves *outsourcing tasks* to a *network of people, also known as the crowd.*		**Mission**: *AgNIC facilitates and participates in partnerships and cooperation among institutions and organizations world-wide* that are committed to the *identification, delivery and preservation of reliable, freely-available, evaluated, digital content and quality servic*es for agriculture, food, and natural resources information.
• distributed problem-solving and production	⟷	• facilitates and participates in partnerships and cooperation among institutions and organizations world-wide
• outsourcing tasks	⟷	• identification, delivery and preservation of reliable, freely-available, evaluated, digital content and quality services
• network of people, also known as the "crowd"	⟷	• AgNIC, [also known as the "partners"]

The key elements of crowdsourcing (distributed, outsourcing tasks, network of people) correlate closely with the expressed mission and organizational structure of AgNIC. For an organization whose culture is based on collaboration, the crowdsourcing of mission-critical tasks was an obvious methodology.

Process and Considerations

Locating and harvesting content for the AgNIC portal began in 2007 when AgNIC successfully piloted the Open Archives Initiative (OAI) protocol for metadata harvesting of five institutional repositories with agricultural content. As the number of IRs with appropriate content increases, AgNIC needed a sustainable way to discover and harvest additional new content. Before the crowdsourcing project began, about 30 repositories were being harvested. To expand this coverage, the Content Committee devised a short-term, manageable and sustainable project that could be accomplished by a cadre of volunteers working from their own locations and at their own convenience.

The name given to the event was "HARVEST," an acronym for Helping AgNIC ReVeal Extraordinary Stored contenT, which seemed appropriate for agricultural librarians. The name was also fitting as we scheduled the event for September, the harvest time of the year. HARVEST was described as "a crowdsourced repository review to identify digital collections containing ag-related content. The metadata from these collections will be harvested and presented as part of the AgNIC resource database subject search."

An announcement about the crowdsourcing project went out to subscribers of the AgNIC listserv. AgNIC routinely uses webinars as a means to communicate with partners throughout the year and between annual meetings. Prior to the scheduled HARVEST, a webinar was held for those interested in participating. It was hoped that by going through the OAI harvesting process, volunteers would have a better overall idea about how content is added to the AgNIC portal and they would be motivated to help find additional content. We covered specifics of the OAI harvesting procedure and why IR review by librarians was important and necessary. The steps in the process for the HARVEST review were demonstrated, followed by time for attendees to ask questions. The webinar was recorded so that anyone who could not attend the webinar could view it before participating in the project. As an added incentive for participation, we announced that there would be prizes given to two randomly selected people who participated in the event. Seventeen people attended the webinar.

Selection and HARVEST Discovery

The AgNIC harvesting process uses the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It is a low-barrier mechanism for repository interoperability, in which data providers (repositories) expose structured metadata via OAI-PMH. The metadata can be harvested for reuse by data consumers, such as AgNIC. The process is straightforward: an AgNIC partner identifies a repository with specific desirable collections; an AgNIC representative emails a contact person at the IR asking permission to harvest portions of their content; the staff obtains the "base url" for the desired collections and the technology support team programs the harvester to run. Periodically, the harvester visits the repositories, retrieves the metadata from the designated collections, normalizes it a bit, and finally adds it to the AgNIC resource database index. The background process runs transparently, regularly updating the index as new resources are added to the targeted repositories.

In order to achieve our goal of adding more content to the AgNIC resource database through the HARVEST, we needed to evaluate and select additional IRs with suitable content and harvest the metadata. It was important that this evaluation step be done by partners as AgNIC is committed to selecting content that meets our quality standards. We created a spreadsheet with institutional repositories, tabbed by platform (D-Space, BePress, etc.), which included around 650 worldwide IRs. For this initial harvest, we focused on the D-Space repositories. At the conclusion of the webinar, we invited the crowd participants to review IRs and select for harvesting those with suitable content over a two-week period. Some people had special language skills which enabled them to look at repositories in languages other than English. In addition to the IRs by platform, we had a list of land-grant repositories and AgNIC partner institutions which were reviewed by a few of the volunteers. These institutions were high priority targets because of the assumed relevant content of any existing repositories.

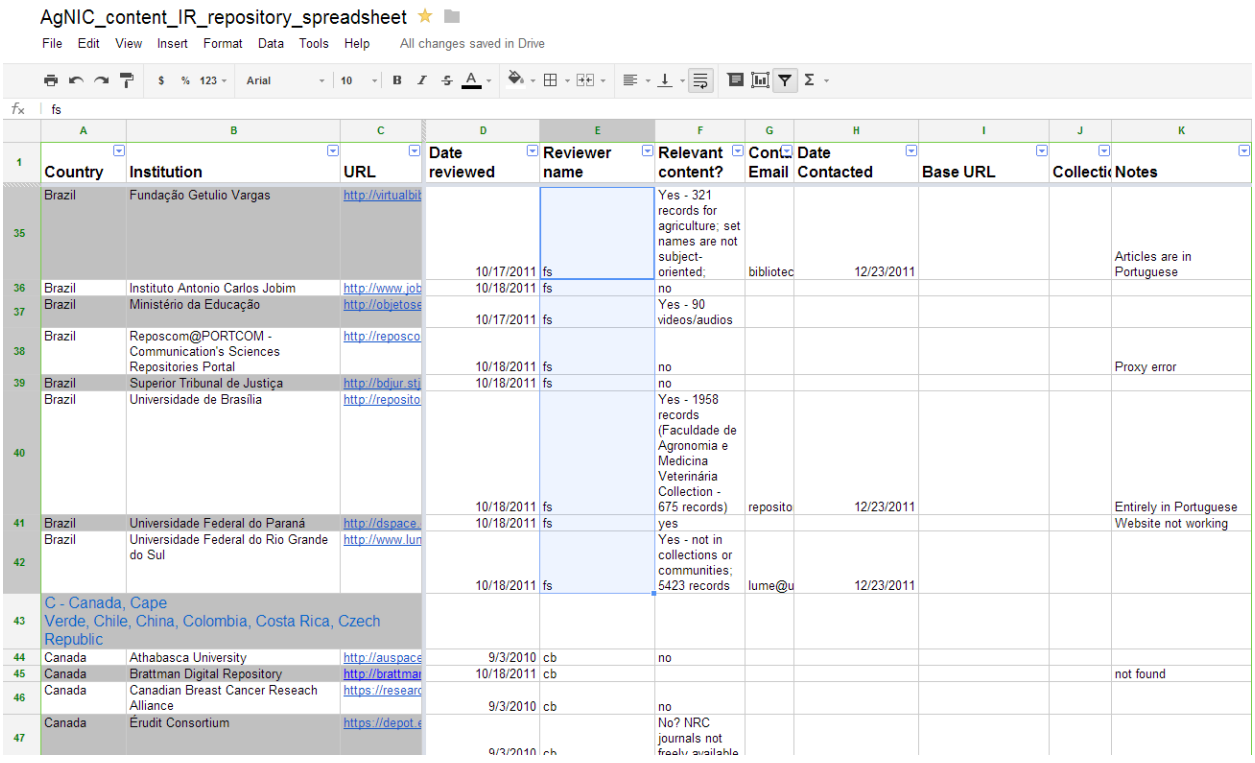
To make the selection and review more consistent, we provided a short list of tips for the crowd.

HARVEST Project Tips:

- Use the URL for the library or IR provided on the D-Space or land-grant list
- Scan the homepage or troll the site for a digital repository or an index page for digital collections

- Browse for communities with “agriculture-like” names
- Use the “search” feature on the site if possible and use agriculture/natural resources keyword
- If you know the university has a strength in agriculture or natural resources, delve into the collections for that subject area
- If you aren’t sure of any agricultural strengths, do a search or two on the IR and try and locate materials on an agricultural topic. If nothing is forthcoming move on…
- If you find only a few relevant documents? Add a note to come back later.

The ability to share a single spreadsheet via Google Docs made it possible for all participants to edit a single document. The Google Docs spreadsheet allowed for up to 50 simultaneous editors, which was ample for our crowd. The spreadsheet contained basic information about the name, geographic location, sponsoring institution, and URL for each institutional repository, grouped on separate tabs by the host platform. Additional columns provided space for the participants to record their name or initials, the date of the review, and a yes/no harvest recommendation, as well as any additional notes. This single spreadsheet enabled us to track the followup steps as the information was acted upon.



AgNIC_content_IR_repository_spreadsheet

File Edit View Insert Format Data Tools Help All changes saved in Drive

	A	B	C	D	E	F	G	H	I	J	K
1	Country	Institution	URL	Date reviewed	Reviewer name	Relevant content?	Cont… Email	Date Contacted	Base URL	Collecti	Notes
35	Brazil	Fundação Getulio Vargas	http://virtualbit	10/17/2011	fs	Yes - 321 records for agriculture; set names are not subject-oriented;	bibliotec	12/23/2011			Articles are in Portuguese
36	Brazil	Instituto Antonio Carlos Jobim	http://www.job	10/18/2011	fs	no					
37	Brazil	Ministério da Educação	http://objetoss	10/17/2011	fs	Yes - 90 videos/audios					
38	Brazil	Reposcom@PORTCOM - Communication's Sciences Repositories Portal	http://reposco	10/18/2011	fs	no					Proxy error
39	Brazil	Superior Tribunal de Justiça	http://bdjur.stj	10/18/2011	fs	no					
40	Brazil	Universidade de Brasília	http://reposito	10/18/2011	fs	Yes - 1958 records (Faculdade de Agronomia e Medicina Veterinária Collection - 675 records)	reposito	12/23/2011			Entirely in Portuguese
41	Brazil	Universidade Federal do Paraná	http://dspace.	10/18/2011	fs	yes					Website not working
42	Brazil	Universidade Federal do Rio Grande do Sul	http://www.lun	10/18/2011	fs	Yes - not in collections or communities; 5423 records	lume@u	12/23/2011			
43	C - Canada, Cape Verde, Chile, China, Colombia, Costa Rica, Czech Republic										
44	Canada	Athabasca University	http://auspace	9/3/2010	cb	no					
45	Canada	Brattman Digital Repository	http://brattma	10/18/2011	cb						not found
46	Canada	Canadian Breast Cancer Reseach Alliance	https://researc	9/3/2010	cb	no					
47	Canada	Érudit Consortium	https://depot.e	9/3/2010	cb	No? NRC journals not freely available					

The instructions to the participants were very simple: during the period designated for the HARVEST and at their convenience, follow these steps:

1. open the Google Docs spreadsheet;
2. view one of the IR listed that has not already been reviewed;

3. correct the IR URL if necessary;
4. enter the date you reviewed the site;
5. enter your initials;
6. indicate (yes, no, maybe) whether there is suitable content to be harvested;
7. enter a contact email address of the IR, if available.

The planned two-week event was extended an additional week to capitalize on the momentum and expand the opportunity for others to participate. A total of 13 partners participated in the Harvest, some in one session, others in many sessions over the three-week period. Ultimately, any level of participation contributed to the overall goal and advanced the project. Thanks to a hardy group of 13 crowdsource volunteers, over the three week event 328 institutional repositories were reviewed for agriculture-related content suitable for harvesting by AgNIC for the resource database. Of these, 146 sites were identified for harvesting. Once these are incorporated into the database, AgNIC will expose many new records to those who use the general AgNIC search.

Next Steps

Identifying the new IRs for harvesting is a significant accomplishment, but more work remains to be done in order to complete the process. Some of this can be accomplished through further crowdsourcing; some steps require the action of technical staff. For each repository, we sought to identify an administrative contact and, as a matter of courtesy, requested permission to harvest the metadata. In most cases, we only want to harvest a portion of the repository content, so we needed to determine the base URL and the identifiers for the specific collections. These elements were given to the AgNIC Secretariat technical support to update the harvester and schedule the harvester to run.

The Committee has identified additional groups of IRs to review, with those using the Fedora platform a priority. Repositories are continually adding new collections and growing, so it will be necessary to periodically reexamine these IRs to ensure that all relevant content is included in the harvester.

Conclusion

For librarians who are used to shared cataloging, distributed collaboration may not seem like a novel activity. However, crowdsourcing our efforts to review a large group of institutional repositories for desirable content was just the incentive needed to move the project forward. By creating a structured opportunity, we were able to spread the work among a number of volunteers, relieving the Content Committee and the AgNIC Secretariat of the full responsibility. Creating an event held over a limited period of time helped focus the project and make it attainable. There was a certain amount of camaraderie that was developed among the participants, who were motivated by the goals, the activity and the prizes. As one participant commented, “It was fun to look at all the different repositories around the world, and a great idea to do this as a crowdsourcing activity.”

Key to the success was having a clearly defined goal, specific steps and directions, a technology solution for tracking the progress and efforts, and a few individuals who assumed the leadership necessary for planning and promoting the activity. We have demonstrated that this crowdsourcing approach works for our group and plan to utilize it to continue the current project and for other initiatives.

Agricultural Network Information Center (2012), "About the AgNIC Partnership", available at: http://agnic.org/about/ (accessed 11 February 2013).