Using Webinator to search online collections of Eurasian and East European research

The Center for Russian & East European Studies, a sub-unit of the larger University Center for International Studies (UCIS) at the University of Pittsburgh, won a competition a number of years ago to create the Vladimir I. Toumanoff Virtual Library — a collection that includes searchable online documents from many top U.S. researchers and analysts who write about politics, history, sociology, economics and foreign policy related to the states of the former Soviet Union and Central and Eastern Europe. Thunderstone's Webinator indexing and retrieval software enabled the responsible Informatics team to accomplish this goal in an efficient and affordable manner.

The University Center for International Studies (UCIS) provides the organizational framework that supports the University of Pittsburgh's mission to integrate and reinforce all its strands of international scholarship in research, teaching and public service. UCIS includes — in addition to many other highly-acclaimed programs and component units — a Center for Russian & East European Studies, an Asian Studies Center, a Center for Latin American Studies, a European Studies Center, an International Business Center (jointly sponsored with the Katz School of Business) and a European Union Center of Excellence (funded by the European Union.)

As a thin layer on top of the whole UCIS structure, Central Administration handles all business-related core functions and technology issues. When individuals in any of the sub-units need advice or consulting related to I.T. Services, Knowledge Management, database planning, upgrading of their websites or anything else that would fall into technology- mediated information, they call upon Mark J. Weixel, Director of Informatics at UCIS.

Discovering Webinator and Getting Started With Using It as an Easily Customizable Development Tool

Weixel recalled, "Back in I guess it was '98, I found out about Webinator from a friend of mine who was at Princeton at the time. We had a particular niche here in International Studies, and we wanted to create mini search engines for web content that was specific to certain world regions. We were hoping to create search engines like AltaVista, since Google wasn't even around then, that would allow people to do full-text searching of those websites. But, because we were vetting the list of sites, we thought we could increase the probability that searchers would come across something really relevant to the part of the world we were focusing on.

"We used Webinator to index and search collections of websites that were in and dealt with Russia and Eastern Europe.

"So, that was my original introduction to Webinator. We bought the entry-level product to begin with, and we currently have the Enterprise version. What I really like about it, still, is the fact that it's relatively easy to configure. It's much easier to configure that it was back when we bought the original product, when everything was run through command lines. I like the notion of relevance in terms of returned hits. It seems to make a lot more sense to me than, for example,

Google page ranking — which places a much higher priority on popularity than it does on the actual content of the pages where text matches.

"Another thing that has been nice is the fact there is support for synonym matching within the server. And I think Vortex as a scripting language is very powerful. Even though I haven't used it to its fullest ability, it's proven to be quite flexible when we've needed to make modifications."

Implementing a Sophisticated Indexing and Retrieval Package with an Attractive ROI Track Record

Did they look at any competing products? According to Weixel, no, they didn't — for a couple of reasons. One, they're a small shop and they have to ask, "How much is this going to cost?" And, he said, the ROI for a one-time investment in a perpetual Webinator license was always pretty clear. It was a known quantity to them. Plus, Weixel strongly believed, as the person in charge of actually setting up and administering it, Webinator provided an affordable and high-quality solution for his specific application requirements. The business manager trusted Weixel's judgment, and by all accounts Webinator has delivered excellent results.

"Vortex as a scripting language is very powerful. It's proven to be quite flexible when we've needed to make modifications."

As to future expansion beyond the Center for Russian & East European Studies, discussions have begun with several of the other sub-units within UCIS. The Center for Latin American Studies and the European Studies Center also seem interested in putting more and more of their materials online — newsletters, conference reports, etc.

Webinator offers UCIS sub-units the possibility of acquiring a well-proven search engine that they could customize as desired and manage on their own.

Digitizing, Capturing and Making Searchable the Publications that Comprise the Vladimir I. Toumanoff Virtual Library

Weixel said their Webinator-powered search implementation getting the heaviest use right now is a project that the University of Pittsburgh's Center for Russian and East European Studies (REES) has done in conjunction with The National Council for Eurasian & East European Research (NCEEER, frequently pronounced 'Nickser') — a federally funded organization charged with supporting research, typically in social sciences, focusing on the former Soviet Union and Eastern Europe.

REES won a competition a number of years ago to create the Vladimir I. Toumanoff Virtual Library comprised of research reports and working papers submitted to NCEEER by scholars under their grants over the last two decades. This collection includes searchable online documents from many top U.S. researchers and analysts who write about politics, history, sociology, economics and foreign policy related to the states of the former Soviet Union and Central and Eastern Europe. NCEEER continues adding to the collection as its funded researchers prepare new papers.

"We proposed scanning and digitizing more than 20 years' worth of reports and then taking it and essentially pointing Webinator at it and, using the documents plug-in, doing a full-text index of the entire corpus. And I think one of the reasons that we won the competition is because, once we had done the really hard work of creating PDFs out of all the printed documents — we were going to be able to put it in once place and, overnight, have a full-text search index. It's my understanding that that was not a

component of the other proposals," said Weixel.

He continued, "We successfully contended for that particular project, got it, spent the better part of nine months digitizing the materials and, I kid you not, it took, I think, less than 24 hours, and we had a fully searchable index of the entire corpus of research products. And it worked out well. We have this nice, targeted archive of material. We've got it set to re-index on a regular schedule, so anytime NCEEER gets a new batch of project reports — they upload them, they get caught in the next cycle of indexing, and it makes us very happy.

"The search interface for the archive materials of NCEEER is available through the Vladimir I. Toumanoff Virtual Library at the website of The National Council for Eurasian & East European Research (http://www.nceeer.org/toumanoff.php.) You kick off the search there, and then you're transported to Pittsburgh for the actual results set.

"I had support from people who could actually help me efficiently accomplish what I needed to do. That worked out very, very well."

"Recently we put the server housing Webinator behind the firewall as part of our new increased security policy at the University of Pittsburgh. The fact that the folks at Thunderstone — John, in particular, in the Support Group — were able to work with me in coming up with a way to take a search query and pipe it through a back door into Webinator and then take the result set and present that to users in an accessible front-end, was just fantastic. It took me about two weeks once I had access to the beta version of the code, and that worked out really well. It was satisfying for me on a number of levels, not just because the product did what it was supposed to, but because I had support from people who could actually help me efficiently accomplish what I needed to do. That worked out very, very well."

Weixel added, "Our audience is interesting. Of course, we're housed within a major research university. So, we do have a number of our projects where we're trying to target our students and our faculty. But the area studies centers, these sub-units underneath the University Center for International Studies, most of them have federal funding that mandates what they call 'outreach' — trying to bring the message of international studies to a larger community, whether it's a local business community or whether it's local educators at the Kindergarten through high school level. Most of them probably have some kind of academic interest in one of the regions of focus. However you look at it, it's a pretty large and diverse audience.

"Being in an international studies environment, one thing that is important to us is foreign language support. I will admit to not having tried this yet with any of the CJK languages. But, in terms of the European and Cyrillic-based languages that we've indexed, Webinator has been a really good performer. And we've been quite happy with that."

For more information about UCIS or any of its area studies centers, you may contact UCIS at:

University of Pittsburgh
University Center for International Studies
4400 Wesley W. Posvar Hall
Pittsburgh, PA 15260
ucis@pitt.edu