Using xSearch for Accelerating Research—Review of Deep Web Technologies Federated Search Service

Abstract

Developed by Deep Web Technologies (DWT) (<http://www.deep-webtech.com/>) in consultation with the Stanford University Libraries and Academic Information Resources, xSearch (<http://xsearch.stanford.edu>) is a locally named and customized version of DWT’s federated search service that can search 170+ resources at one time. (Explorit is the commercial version of Deep Web Technology’s federated search solution which Stanford has locally branded as xSearch.) In addition to the broad range of resources searched, xSearch provides an intuitive interface for searching, refining, and displaying results. The ability to choose the resources searched allows this tool to be used by researchers as well as undergraduates. Speed is enhanced by the fact that some results are displayed before the search is completed. Users are able to easily set up alerts from their search results. Both library staff and users are able to create custom search pages that can be a link on a Web page or embedded as a search box in a research or course guide. xSearch has provided a scalable, cost-effective, and robust solution for meeting a wide variety of information needs.

Pricing Options

Customized pricing is provided for each agreement. DWT’s pricing structure for Stanford contains: (1) Fixed and initial startup costs which include a software license for the Explorit Federated Search application, custom development, and support and maintenance fees for one Standalone Implementation, as well as initial connector development for resources. (2) Annual recurring fees to maintain connectors and for Explorit maintenance and support. (3) Capability to add new connectors, which will incur an additional one-time fee for setup and development, and a lower annual recurring fee to maintain the connector. We found DWT’s pricing to be competitive and, in some cases, significantly lower than similar products available on the market. All of the features developed for Stanford are available to other DWT customers. DWT also offers additional features not currently included in xSearch (e.g., multilingual searching, multimedia and image searching).

Explorit, DWT’s federated search engine, implemented at a mid-sized academic university with a population of up to 25,000 students and federated search services that contain 100 resources on a server hosted at DWT would cost about $20,000 a year. This price includes alerts services and the ability to build custom search engines.

Product Description

DWT has developed several federated searching sites that may be familiar to users. Sites include: Science.gov (<http://www.science.gov>), WorldWideScience.org (<http://worldwidescience.org>), SciTopia (<http://www.scitopia.org>), ScienceResearch.com (<http://www.scienceresearch.com>), Mednar (<http://mednar.com/mednar>), and Biznar (<http://biznar.com/biznar>). xSearch is the name we decided to use for DWT’s federated search services that were customized for Stanford.

CONTENT

Because the intended audience for xSearch (<http://xsearch.stanford.edu>) includes researchers as well as undergraduates, the ability for subject specialists to select the resources to be searched via xSearch is a very important feature. It is worth noting that because content is queried in real time, not pre-indexed, the results include the latest information available from a resource.

Resources in xSearch provide depth and breadth for a wide array of subject areas and document types. In addition to abstracting and indexing services covering articles, conference proceedings, dissertations, patents, and technical reports, it was a high priority to include sites that have full-text searching of ebooks and ejournals in xSearch. Full-text search sites for newspapers and primary historical resources are also included. Other resources selected for xSearch include archival finding aids, government publications, and reference materials. Statistical resources searchable via bibliographic data are also included in xSearch. A few foreign language resources are included, but there are no special features implemented to assist in searching them.
A few of the resources that we wanted to include in xSearch would not work. There were several reasons: design of the search interface (e.g. must browse to select search terms), poor response time for a resource, and lack of persistent URLs for retrieved items. The native interface is being used for many queries but a few resources require using a Web Services API. As a result, some resources are accessible via xSearch that are not available from other federated search services.

SEARCH INTERFACE
End-user access to xSearch (<http://xsearch.stanford.edu>) requires only a Web browser and an internet connection. When working with DWT to customize the interface, great care was taken to make navigation and access ADA-compliant. A mobile version does not exist although many features work on an iPad. So that help information could be modified easily, a “companion” Web page was set up locally (<http://lib.stanford.edu/xsearch>) that contains a list of resources, a quick reference guide and tutorial, plus links to custom search engines created for subjects or courses.

The Quick Search page allows users to search Keywords in a fixed subset of 50 popular resources (see FIGURE 1). The Advanced Search page (see FIGURE 2) allows users to enter terms in Keywords, Title, and Author fields along with date ranges for all 170+ resources. Users may select resources they want to search by choosing individual titles or by choosing categories. Brief descriptions of each resource are displayed when a user hovers their computer mouse over the resource name. A separate search page for searching selected business resources by Company Name is in development.

All search pages support Boolean logic (AND, OR, NOT) plus nesting if users want to combine terms using AND and OR in the same query. Two trun-
users refine results, clusters consisting of Topics, Authors, Sources, and Dates are generated dynamically. Terms in the Topics cluster are semantically generated. Results can be sorted by Rank, Date, Title, and Author. Users can also filter or limit the display of search results to a particular resource (see FIGURE 3). Session Preferences enable users to increase the number of items displayed per page from 15 up to 250.

Once search results are available, there are several features to help searchers identify and export records of interest. Stars are present to indicate relevancy ranking. Several sentences from an abstract or snippet showing hit term highlighting are also included. SFX links or full-text icons are present in the brief record display format. Full record displays for an item take a user into the native interface of a resource. Users are able to email or print entire answer sets from the main menu. If only some items are of interest in search results, users can check a box next to an item and then go to My Selections where they are able to directly export to RefWorks; export to EndNote Web; save in RIS format for importing into other reference manager software; print the selected citations; or email them.

A Results Summary link provides the total number of items found in a particular resource and the total number of items included in xSearch results (see FIGURE 4). A red X appears by the name of the resource if the connector failed to do a search and a stopwatch appears by the name of a resource if a search timed out due to slow response. This information allows users to verify that all the resources that they choose are included in the search results. The Results Summary also helps users identify specific resources they might want to search individually if they are doing in-depth research.

alerts and custom search pages

Setting up alerts in xSearch is easy and intuitive. The search results page includes a link to “create an alert from this search.” Users can easily modify what resources are searched in their alert, select the frequency of alerts (weekly, monthly, quarterly, annually) and their format (email, RSS feed, Atom feed). Alert details (search strategy, resources searched, etc.) as well as results from several updates can be viewed in the xSearch interface by clicking on the My Alerts link.

Both library staff and users can Create a Custom Search (see FIGURE 5). This feature allows users to create a search page that contains only their favorite resources. This service also enables librarians to instantly create custom searches that are tailored to a particular

 viewing search results

While all of the content for a chosen resource is searched, we need to limit the number of items retrieved from each resource so that we do not violate our terms of use agreements with resource providers. That limit varies from 100 to 200 for each resource. Retrieved items are merged into one relevancy ranked list and de-duplicated. It is possible for DWT to set priorities for which version of a duplicate record is displayed.

Because initially retrieved items are displayed while the search is still running, users are able to begin viewing items right away. A colored progress bar provides a visual cue about the number of resources searched. Users are prompted to accept remaining results rather than having them added automatically. This allows users to read through results without the list changing as more items are retrieved. To help
class or research guide. Saved custom searches are accessible through the xSearch interface, as a link that can be added to Web page, or as an embedded search box via widget code.

TECHNICAL INFRASTRUCTURE AND TECHNICAL SUPPORT

Stanford has worked with DWT for four years. Initially, DWT created three federated search prototypes for us: (1) a collection of library catalogs, (2) a collection of locally digitized materials, and (3) a collection of licensed resources that included databases and full-text ebook and ejournal sites. With shelving space limited on campus, before items were shipped off-site faculty wanted to have a more robust search interface for discovering library materials than was possible via Sirsi or DWT. Blacklight (<http://projectblacklight.org>), an open source discovery application, was selected and local development work followed. The new interface to the library catalog is called SearchWorks (<http://searchworks.stanford.edu>). Because SearchWorks also includes bibliographic access to locally digitized materials and because of the considerable investment the library makes in licensing resources that are not available via the open Web, we decided to focus our efforts with DWT on licensed resources. Our efforts included working with DWT in order to get a summary of search results, alerting services, a custom search engine, and support for open URL linking. DWT also customized the search interface for us in order to harmonize the user interface with other library Web applications. While there are no technical barriers to prevent the library catalog from being merged into xSearch, because development efforts have been underway for both interfaces, we have not yet integrated SearchWorks and xSearch.

DWT offers both hosted and locally installed versions of their federated search software; Stanford opted for local installation on its server farm. The initial production instance of xSearch was launched two years ago and included 28 resources running on a RedHat virtual machine. A year ago, when we decided to significantly expand the number of resources, to insure adequate bandwidth for production use and to provide a test environment for future developments, Explorit Search Engine and accompanying software were installed on dedicated dual CPU server with 4 GB of RAM.

Staff support from both DWT and Stanford is needed for xSearch. In addition to technical support provided by DWT staff for custom development, installation, and maintenance, technical support was also needed by Stanford for implementation (installation and configuration of the application, integration with University authentication systems, specification of open URL linking) and for maintenance (ongoing server management, coordination with DWT for upgrades and changes to connectors). Looking ahead, the goal is to have someone at Stanford assigned to xSearch as the Application Administrator. After an initial period of
training and knowledge transfer, it is estimated that this work will consume less than 10% FTE on an ongoing basis for routine monitoring and maintenance. The Application Administrator would be expected to increase their time commitment during upgrades.

**USAGE STATISTICS**

Usage statistics far surpass what is in a typical COUNTER Report. Statistics that are provided include information such as the number of queries, response time for each resource, percent results that are relevant, and the number of click-throughs to full-text. Details for terms searched and particular articles viewed are also available. Daily, weekly, and monthly use statistics are available graphically and can be exported into Excel. Statistics to provide details about alerts were developed at Stanford and include total number of alerts and email addresses for those receiving alerts.

**Critical Evaluation**

Compared to other federated search products, Stanford found that DWT offered the most compelling package of performance, features, and design. While federated search engines’ performance is inherent-ly limited by the performance of its target sites, DWT’s progressive delivery of results gives researchers near real-time response with the first set of responses while the application assembles a complete set of hits from all sources. When we chose DWT, other federated search services we considered (e.g., MetaLib) were much slower because they did not return any results until a search was completed. The recommended number of resources to search at one time through other services was much smaller too. DWT offered many options for customizing the interface, the only service that included alerts, and the only service that allowed us to create customized “search engines” locally.

DWT calculates relevance across search results on the fly, and clusters hits based on predefined criteria, giving searchers the ability to filter results to items from a particular resource is great. The ability to create custom search pages and set up alerts are features not offered by other federated search providers.

**Pricing:**

Deep Web Technologies federated search services are competitive with and in some cases considerably less expensive than competitor products.

**Contract Options:**

Contract language and options were reasonable.
Relevancy ranking of the search engine is robust but results brought into xSearch from each resource are dependent on default sorting order for that resource. If the default sorting order for an individual resource is by date rather than by relevancy ranking, items retrieved and merged in xSearch results will not be the most relevant items from that individual resource. Search results include both ranked and unranked items with unranked items having only some of the search terms present. This means if results are sorted by date then less relevant items may be displayed first.

The Brief Display includes the title, author, abstract, publication title, publication year, and the resource that an item was retrieved from, as well as a link to the full-text. Ideally, entries for articles would also include volumes and pages but this is a minor annoyance as the full citation is available either when the full record is displayed or when a citation is exported to a reference manager.

**Contract Provisions**

DWT is a commercial service that offers federated search solutions to a wide array of customers. Each contract is customized and is based on such factors as the type of organization, the number of resources included in the service, how the service is deployed (hosted vs. stand-alone), the amount of customization desired, support services needed, etc.

Because nearly all resources included in xSearch are currently licensed by Stanford, access to xSearch is limited to students, faculty, and staff with a current Stanford University Network ID. The DWT contract for xSearch defines “deliverables” that include features, functionality, and services. One key component of this license agreement is defining turn-around time for responding to service requests, for example, a prompt response if a connector is not returning results but batching requests for cosmetic changes to the interface. Statistics provided are not COUNTER Compliant but do offer a much richer array of data and reports that include data such as response time for each resource, number of queries, and number of full-text click-throughs.

**Authentication**

Anyone can view the search interface but only current Stanford users can run a search. Because most resources in xSearch are licensed by Stanford, IP filtering and EZProxy authentication using a Stanford University Network ID (SUNet ID) are invoked when a user presses the Search button, even if a user is located on-campus. SUNet ID entries are checked against the Stanford Directory to insure that users are current students, faculty, or staff (also called LDAP). Creating a custom search engine or setting up an alert also requires SUNet authentication.

Before custom search engines were developed, we had users authenticate when they entered the URL or clicked on the link to use xSearch. After custom engines were developed, because these custom engines can be embedded as a search box, the authentication prompt was moved so that it appears when a user presses the search button.

**Author’s References**


DOI URL: <http://dx.doi.org/10.5260/chara.12.4.11>


DOI URL: <http://dx.doi.org/10.5260/chara.12.1.5>

**About the Authors**

Grace Baysinger is the Head Librarian and Bibliographer of the Swain Chemistry and Chemical Engineering Library at Stanford University. She was the project lead for implementing xSearch at Stanford. Grace’s professional interests include end-user searching, collection development and management, and scholarly communication issues. She serves on several advisory boards and committees for database producers and publishers in chemistry. After receiving her library degree in 1981 from the University of Michigan, Grace held professional positions at UM’s Natural Sciences and Chemistry Libraries. In 1989, she accepted the position as Head of the Swain Library. She also served as the Head of Stanford’s Science and Engineering Libraries from 1999-2005.

Tom Cramer is the Chief Technology Strategist and Associate Director of Digital Library Systems and Services for the Stanford University Libraries. In this role, he oversees the technical development and delivery of the full complement of Stanford’s digital library activities, including the digitization, description, discovery, delivery, preservation and management of digital resources that support teaching, learning and research. Tom is one of the founders of the Hydra Project <http://projecthydra.org/> and an active contributor to Blacklight <http://projectblacklight.org/>, open source projects rooted in higher education that deal with digital asset management and discovery. He currently serves as the director of PASIG, the Preservation and Archiving Special Interest Group, and as a steering group member for Open Repositories.