Executive Summary

NRAO has been investigating various approaches to building web applications to search data collections. In the summer of 2009, we started developing search applications using Blacklight and Solr, both open source, which provide a solution stack for online search which allows in-house development to focus on in-house problems. With one software engineer, we developed a highly reused code base and rapidly built demonstrations for 5 separate data collections by the end of 2009. In early 2010, we deployed two search applications using Blacklight & Solr: a collection of theses and a collection of conference proceedings. Early response has been very positive. By late 2010, NRAO expects to have 3 to 5 search applications online built on top of Blacklight, still with one software engineer.

Motivation

So you have a data collection and you’d like to provide an online search tool. How do you get started? At NRAO, particularly in the library and other public-facing divisions, we’ve been facing this challenge. The library set forth this year to put two databases online: NRAO Theses and ISSTT proceedings.

Data Collections

NRAO Theses database. NRAO library maintains a collection of bibliographic records relating to graduate theses written using NRAO instruments. This database has been migrated several times in the past 4 years. It is currently managed in Microsoft Access.

ISSTT Proceedings. NRAO library manages the ISSTT Proceedings, a collection of the papers from all proceedings of the International Symposium on Space Terahertz Technology. NRAO is digitizing and hosting all of proceedings to date, totaling more than 20 years of meetings. The proceedings are a collection of scanned-in PDFs, with metadata managed in MySQL.

NRAO Papers. NRAO library maintains a collection of bibliographic records for publications which use data from NRAO telescopes or are written by NRAO staff, scientific or technical. NRAO has been developing a public search interface in PHP and a separate administrative interface for library-use only. This database is managed in MySQL.

Telescope Proposals. NRAO operates public instruments and has 3 calls for proposals throughout the year. Once a proposal is accepted, its coversheet is circulated to telescope and observatory operations as needed. The Proposal Submission Tool uses MySQL, but this only includes recent proposals. NRAO
has no central proposal database for historical proposals, and currently there is no public portal to access proposals online.

Telescope Data. When data is collected on NRAO telescopes, observers have a proprietary period (usually 12 months) to process and publish results. When the proprietary period ends, data is public and are often useful for archival research. Since 2003, NRAO has been developing an online search and access tool for both public and proprietary data, given authentication. With 3 currently active telescopes (EVLA, VLBA, GBT) and 1 under development (ALMA), the current search tool only includes EVLA (and VLA) and VLBA. NRAO is currently putting GBT data online and working to improve cross-telescope access. The data itself is a collection of flat files. Metadata about VLA, EVLA, and VLBA are managed in Oracle. Metadata about GBT is managed in PostgreSQL.

Other. NRAO is also considering search tools for various other collections, such as technical memos and meeting minutes.

Blacklight & Solr in the Wild

First, let’s dive into a few deployed Blacklight & Solr instances. You can access these here at the meeting as well as at your home institution. We’ll get into the details later.

Stanford University has a Blacklight OPAC at http://searchworks.stanford.edu/. This is a great example of what Blacklight is intended to do: Online Public Access Catalog.
NRAO has been developing with Blacklight since the summer of 2009 and launched two applications in early 2010. We are not using Blacklight for a library catalog, but as an interface to specific data collections. The first instance is the NRAO Theses database at http://find.nrao.edu/theses.

The second NRAO Blacklight instance is for ISSTT proceedings at http://search.space-thz.org.
Solr - [http://lucene.apache.org/solr](http://lucene.apache.org/solr)

Solr is an open source product available from the Apache Software Foundation (ASF). Apache is best-known for developing the most widely used web server called Apache httpd, which most people simply call Apache. Today, the ASF provides much, much more. Solr is built on a developer library called Lucene. Lucene is a Java library for adding powerful search indexing to your application. Solr is a stand-alone search server, where developers need no knowledge of Lucene itself for most use cases. Customizing Solr is almost always just a matter of managing XML configuration files.

Blacklight - [http://projectblacklight.org](http://projectblacklight.org)

Blacklight is an open source Ruby on Rails project which provides a web user interface to Solr. The project has an actively growing community, led by libraries at the University of Virginia and Stanford University. Blacklight is built for library catalogs and works with MARC records out of the box. If you are not using MARC, you can configure Blacklight using a simple configuration file. Customizing the Blacklight user interface (UI) is just like customizing any Ruby on Rails project. Simply editing the Blacklight configuration file and branding the UI will go a long way to building your own app.

**Developing Your Own Blacklight**

In our experience at NRAO, developing a custom Blacklight instance has three phases.

**First, define your data.** Using Blacklight & Solr to search your data is much easier if your data is well-defined. In fact, building an index in Solr will help you understand your data. Spend some time defining the fields you’d like to search and which fields you’d like to display. Keep in mind that using Solr means you are denormalizing your database. If you have a relational database, you are in effect packing all of your fields into a single table.

If you’re ready early on, you can follow a Solr tutorial and start with a very simple schema. If you are the computer geek type, this is likely easier than you think. The first time we looked at the theses database, we were able to demonstrate our search index in Solr in less than 15 minutes.

**Second, install and configure Solr and Blacklight.** Blacklight provides a demo application which is a good starting point. Once you are comfortable with the demo application, set up development instances of Solr and Blacklight (instructions are online). Configure both to match the schema of the data you defined in the first step. Both the Solr and Blacklight communities are very supportive and will help if you have questions or run into issues.

**Finally, customize the Blacklight UI.** Change the colors and logos to those of your institution. Adjust fonts, page widths, and other style elements. Add links to useful resources, keying off fields in a given document. Get creative.

**Repeat.** The faster you iterate the better. At NRAO, we developed new features daily, thanks to the great features and development hooks in Blacklight & Solr.