Katipo Communications Ltd
PO Box 12 487, Wellington, New Zealand
Ph: +64 4 383 5526 | Fax: +64 4 934 1286
www.katipo.co.nz | walter@katipo.co.nz

April 21, 2008

Horowhenua Library Trust
Attn: Joann Ransom

Katipo Report

to Horowhenua Library Trust
re Federated Search Report

Federated Search Report

Introduction

Horowhenua Library Trust, referred to as HLT hereafter, is interested in including results from outside of the Kete Horowhenua site in the site's searches. Katipo has prepared this report on possible strategies HLT could take to achieve this result.

The sample outside data sources that HLT has outlined fall into two categories; those that HLT runs itself and those run by external organisations.

HLT's data sources include the content of the Koha library system, HLT Cemetery Records, and the HLT Clubs and Organisations database. All of these applications use a MySQL relational database to hold their information. Since they are operated by HLT itself, access to the databases can be arranged without negotiating with another organisation.

HLT would like to also search records from National Library of NZ Timeframes database and the Online dictionary of NZ biographies. These are maintained by external organisations and rights to use their data must be negotiated.

HLT is also interested in exploring other potential sources of external data. Some of these hypothetical source organisations may have blanket policies about how another site may use their data and where the site may access it or they may need to be negotiated on a case by case basis.

In this report, Katipo will describe HLT and the Kete software's technical assets, the issues that searching disparate source databases from Kete Horowhenua may raise using the mentioned databases as examples, and strategies that HLT might use to incorporate this outside data.

HLT And Kete Software's Assets

Kete Assets

Since Kete Horowhenua is the ultimate target where federated search for HLT will be implemented, we must consider what it already has so that we can evaluate strategies and whether its functionality needs to extended.

Kete has the following assets that relate to federated search:

1. an existing User Interface that handles searching different record types (topics, images, etc.)

2. an instance of a Z39.50 search index database implemented using the Zebra application that can be queried by Kete or by other applications

3. use of the OAI-PMH Dublin Core standard schema to populate its Z39.50 search index database and return search result records

4. conventions that utilize OAI-PMH Dublin Core to suit Kete's handling of search results, such as the “oai identifier” XML field containing the host name of the source data, the directory path it is contained in, the record type, and its internal id number (e.g horowhenua.kete.net.nz:site:Topic:1)

5. Zebra's included software tools for indexing data records separate from the Kete web server, including use of XSLT for transforming a record's format

6. existing Kete code, including the Ruby on Rails framework it is based on, for formatting, outputting, or digesting data

7. a plan by Katipo to implement an included (though optional) OAI-PMH Repository in the Kete software funded by the Aotearoa People's Network and using the Ruby OAI utility library (http://oai.rubyforge.org/)

HLT Assets

Beyond Kete, HLT has three databases that it would like to search from Kete Horowhenua. All three are implemented using a relational database behind a web application interface. The relational databases are quite capable of being accessed by other software applications. In our case it would be in a “read only” manner.

HLT is also a part of larger group of Kete users that will be extending the software, sharing knowledge, and possibly providing external data that HLT may search.

Issues

HLT's Databases

All three databases have different data structure conventions from Kete and different again from each other. Their existing data models will need to be translated to something that Kete understands in order to be represented in Kete searches. Since Kete uses OAI-PMH Dublin Core in its own search results, this is the most logical target for translation.

None of the three existing applications have a “machine readable” repository of their data, except in the form of direct access to their relational databases. The Koha library system can, in later versions, provide a Z39.50 protocol interface, but HLT uses Koha 2.2.2 and does not currently have access to this functionality.

External Data Sources

Katipo explored options for both NZ Timeframes and the Dictionary of New Zealand Biography. No existing automated standardized interface for federated searching agents to use was found for either.

Enquiries were sent to both organisations asking about available ways of extracting data from their sites. NZ Timeframes did not answer, but Dictionary of New Zealand Biography (DNZB) responded that their long term plans are to move to the platform that Te Ara uses. Te Ara currently has standard SRU/W access and will likely add an OAI-PMH Repository. DNZB were very interested in being a test case for project's federated searching ambitions.

Although NZ Timeframes doesn't provide an outward facing facility for OAI-PMH Dublin Core that Katipo could find, it is known that they supply a subset of their records in that format to the Matapihi federated search site. This is most likely a periodic manual process. Perhaps NZ Timeframes could be convinced to share these records with Kete Horowhenua, too?

It is possible to do what is called "screenscraping" to harvest results in HTML from these sites' web search facilities either at periodic times and load them into Kete Horowhenua's search index database or at the time of end user query on Kete Horowhenua, however screenscraping is usually labour intensive to develop and prone to break over time since there are no agreed upon standards with the data source. Katipo wouldn't suggest this technique in these circumstances since in the long term there will likely be alternatives.

So for these two potential data sources there is no immediate way forward, though use of techniques around the OAI-PMH protocol sound like the most promising in the long term. Especially in consideration of Kete's existing and future functionality.

It should be noted that other online resources do provide functionality to be accessed by outside applications. For example, http://nzresearch.org.nz/ provides an extensive OAI-PMH repository with sets covering 15 educational institutions.

A number of New Zealand libraries provide Z39.50 access to their catalogs (http://wiki.lianza.org.nz/index.php/Resources/Z3950ConnectionInformation). A quick survey found that most didn't include a pre-defined URI for an online web catalog record and were in MARC bibliographic format. Thus work would be necessary to figure out a record's URI in each library's online catalog, probably on a library by library basis, and also to parse the MARC data.

These Z39.50 do provide a service that can be accessed at the time of the Kete user's search rather than holding (and duplicating) the metadata for records in the Kete instance's own search index database, but there would be computational overhead in setting up the network connection for each Z39.50 source, parsing the results into a schema Kete understands, and sorting the results with other sources. There may also be interruptions to certain sources associated with a connection over the network. With this model, the more sources, the more connection and sorting overhead is necessary.

HLT has also expressed interest in searching other Kete software instances. Kete can provide a Z39.50 interface to it's records' metadata via the Zebra application. Soon Kete will also provide an OAI-PMH Repository interface, too.

Recommended Strategies For Kete Horowhenua

Katipo recommends that HLT consider implementing the following:

1. create a simple open source application that translates non-Kete systems' databases to an OAI-PMH Repository, with support for relational databases as a read only data source

2. add OAI-PMH Harvesting to Kete with an administrator web interface to configure and schedule harvests

3. adjust Kete's underlying search code to handle non-Kete types of records (i.e. a record about a book vs a topic)

4. add to and refine Kete's search User Interface to reflect the availability of outside sources

Point 1 would be aimed at making HLT's Koha, Cemetery, and Clubs and Organisations databases available to be harvested by Kete Horowhenua after point 2 is implemented. However, by making the OAI-PMH Repository application for non-Kete systems open source and generalized, HLT would be helping to enable other institutions to make their records available in a protocol that Kete uses and thus encourage more potential data for future Kete Horowhenua searches.

Point 2 will also allow for scheduled harvests from other Kete instances that choose to make themselves available as a OAI-PMH Repository after that work has been completed for Aotearoa People's Network. Bringing other sources metadata about their records into a Kete instance's Zebra search index database should help speed, scalability, and reliability. The downside is that Kete search will not reflect up to the minute changes in the source database.

Points 3 and 4 are necessary to make newly included data from outside sources functional and comprehensible for Kete Horowhenua's users.

For HLT's databases, all these steps are necessary. If HLT decides to limit Kete Horowhenua's federated searching capabilities to external databases that already provide an OAI-PMH Repository (future versions of Kete, for example), than point 1 may be skipped, but all other work is still necessary.